Of Mice and Madmen:
The Other Problem of Radical Indeterminacy

Edward J. R. Elliott

School of Philosophy, Religion and History of Science
University of Leeds

Abstract

Anti-individualistic varieties of functionalism focus not on the causal role a state plays within each individual, but rather on the role a state typically plays in members of that individual's kind. An overlooked problem of indeterminacy arises for anti-individualistic functionalism when there are too many physical states with which mental states can be identified. Taking Lewis' analytic functionalism as my main stalking horse, I show that the (unmodified) Lewisean position implies under reasonable assumptions that for each individual and any system of mental states, there is a maximally 'fitting' scheme of interpretation according to which that individual has those mental states. The result, in other words, is not just one of *radical* indeterminacy, but potentially even *maximal* indeterminacy: anyone can be interpreted as believing and desiring anything, as being in pain or not in pain, and so on. Any solution to this problem requires altering the constraints on 'fit'. I discuss several potential solutions open to the Lewisean, including (i) a new potential role for naturalness in Lewis' theory, and (ii) an alternative approach to understanding of the kind of 'typicality' most relevant to folk psychology.

1. Introduction

Many philosophers will be familiar with the classic indeterminacy worries that often arise in connection for broadly functionalist theories of belief and desire. They usually go something like this: we start with some agent whose behavioural dispositions fit nicely with some reasonable interpretation of their beliefs and desires; we then twist the assigned beliefs in one way whilst twisting the assigned desires in a countervailing way, so as to end up with a very different (and usually quite odd) system of beliefs and desires that nevertheless fits with all the same behaviours and dispositions. We now have two radically distinct assignments of attitudes that are functionally isomorphic in terms of how they causally relate to behaviour, and consequently we cannot use the facts about behaviour to tease them apart. The upshot is we must find some further non-behavioural factor or factors to help pin down which beliefs and desires our agent *really* has, or else face the consequence that our beliefs and desires are radically indeterminate.

Problems along these lines have been discussed by Stalnaker (1984: 15–20), Lewis (1983: 373–7, 1986: 38–9) and more recently Williams (2016, 2018, 2019: 17–32), among many others. Stalnaker proposes to solve the problem by supplementing the purely 'forwards-looking' functional roles that connect systems of belief and desire to behavioural dispositions with further 'backwards-looking' roles that connect the contents of our beliefs to the external world via sensory inputs. Lewis, by contrast, argues that *even if* take into account both the forwards-looking roles in connection with behaviour and the backwards-looking roles in connection with evidence and perception, *still* this won't be enough to pin down a determinate enough assignment of beliefs and desires. His conclusion is that we need further constraints independent of functional role—a bias towards favouring belief-desire interpretations with contents that are overall more natural and reasonable than their alternatives. Along similar lines, Williams argues that the best solution to such problems will require constraints of 'substantive' (as opposed to 'structural') rationality on the assignment of belief-desire interpretations.

So it's an important problem that any functionalist theory needs to deal with, and one that's received a good deal of attention. We might call it the *problem of functionally isomorphic roles*. But that's not the problem that we'll be talking about—we'll be talking about the *other* problem of indeterminacy. Our problem arises not when there are radically distinct systems of mental states that fit equally well with a given pattern of sensory inputs and behavioural outputs, but instead when there are too many ways of typing the physical states of a subject with which those mental states are supposed to be identified. We'll refer to it as the *problem of functionally isomorphic occupants*.

Rather than discuss this problem in way that's neutral between the several different varieties of functionalism, I will instead focus on how it arises specifically for a Lewis-style analytic functionalism and how one sympathetic to that position might want to respond. But the key characteristic of the Lewisean theory that gives rise to the problem is that it is anti-individualistic: mental states are identified with physical states not on the basis of how the latter behave within any particular individual, but instead on the basis of how they typically behave across many individuals belonging to a general kind. I will argue that the unmodified Lewisean view has the consequence that for *each* individual and *any* system of mental states, there will likely be an assignment of those states to that individual which 'fits' the typical facts about the relevant kind at least as well as any other. So we're not just getting *radical* indeterminacy, in other words, but *maximal* indeterminacy.

The essay begins with some background and ground-clearing: §2 clarifies several importantly different ways that indeterminacy can arise within a functionalist theory, and §3 then clarifies some matters in connection with the statistical notion of 'typicality' that's central to the Lewisean theory. Following that, §4 and §5 discuss how the problem of functionally isomorphic occupiers arises for the original, unmodified Lewisean theory. Finally, §6 discusses some ways of modifying that theory to avoid the problem, including (i) adding a new role for naturalness, and (ii) adopting an alternative, non-statistical understanding of the kind of 'typicality' that's most relevant to folk psychology.

2. The Sources of Indeterminacy

Folk psychology includes two kinds of hypotheses. First, folk psychology purports to tell us what kinds of mental states exist; and second, folk psychology purports to tell us how those mental states do and do not relate to one another and to the rest of the world via, e.g., sensory inputs and behavioural outputs. An extremely oversimplified statement of the belief-desire part of folk psychology might say, for example, that there are such things as beliefs and desires, that these are relations to proposition, that an agent can have beliefs and desires towards any propositions whatsoever, and that agents will tend to behave as would maximise how many of the propositions they desire end up being true if all the propositions they believe happened to be true. The theory thus tells us what kinds of belief and desire states are possible, and for each such state it specifies a role the state is supposed to play that connects it to other mental states and to the outside world.

We do not assume that folk psychology includes the whole truth and nothing but the truth of psychology, but it does have plenty to say and it's not unreasonable to suppose that it comes close to the truth. Close enough, at least, that we might use it to implicitly define the terms of our mental vocabulary. This is the thesis of Lewis' analytic functionalism: that 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. Assuming (as we will henceforth) that physicalism is true, the idea is therefore to match up the physical states that agents might be in to the mental states folk psychology posits such that the former *occupy* the roles of the latter—i.e., such that the way those physical states relate to one another and to the rest of the world will mirror the way their associated mental states are supposed to relate to one another and to the rest of the world according to folk psychology. On the oversimplified conception of folk psychology just described, this will roughly mean trying to identify states of belief and desire with physical states such that agents in those physical states will tend to behave as would maximise the satisfaction of those desires if those beliefs were true.

It'll be helpful if we formulate all this in terms of *schemes of interpretation* (cf. Lewis 1983: 119–20). A scheme of interpretation tells us, for each agent of a given kind, if and how any physical state that agent might be in is to be associated with a mental state and vice versa—if you're in this physical state, then you're in that mental state; and if you're in that mental state, then you're in this physical state. The precise details won't really matter for what's to follow, but if we wanted to make that basic idea more precise we could do so as follows. First, take the set of all physical states $P_1, P_2, \ldots$ an agent might potentially be in, and the set of all mental states $M_1, M_2, \ldots$ an agent might potentially be in according to folk psychology. We then let a scheme of interpretation, i, be a partial surjective or non-surjective function from the former to the latter, where we read $i(P) = M$ as saying, of any agent belonging to the relevant kind, that if that agent is (or were) in P then they are (or would be) in M, and vice versa. (Why 'partial'? Because not every physical state need be identified with a mental state; likely most won't be. Why 'surjective or non-surjective'? Because we can't assume that we'll find an occupant for every role that folk psychology posits.)

We should like a scheme of interpretation that aligns nicely with folk psychology. At first pass, we can say that a scheme *fits* to the extent that each physical state occupies the functional role of the mental state, if any, with which it's associated.[1] We say that the scheme i is *perfect* when the function is surjective and $i(P) = M$ iff P perfectly occupies the role of M. A perfect scheme exists only if folk psychology is entirely true; very likely we will need to settle for some imperfection. Finally, we say that the *correct* scheme is the most fitting scheme, if there is just one, and if there are multiple schemes tied for best fit then we say the truth is indeterminate between them (Lewis 1983).

Indeterminacy by itself is not inherently problematic. It would be surprising, in fact, if there weren't a bit of indeterminacy some of our beliefs and desires, and there may even be rare cases of actual agents with very highly indeterminate attitudes. We have a *problem of indeterminacy*, on the other hand, if we find ourselves having to say that the beliefs and desires of all or most actual human agents are radically indeterminate. That's the kind of result we need to avoid. (Or so I'm going to assume.) But there are several ways that a problem of indeterminacy might potentially arise, and we should be careful not to conflate them.

The first and most obvious way in which indeterminacy might arise would be if there are no perfect schemes of interpretation, and multiple imperfect schemes that can all be thought of as having best fit. This may occur if there's no precise fact of the matter about how we ought to measure *overall* fit, and different ways of drawing compromises might give different results regarding which schemes should be considered best. Or it may be that we've settled on a precise method for measuring overall fit, but more than one scheme maximises 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. Lewis was well aware of this 'indeterminacy of compromise', and it did not much bother him—he did not believe it was likely to give rise to *radical* indeterminacy, and so he ignored it (see Lewis 1974: 343, 2020: Letter 450 [1979]).

A rather different source of indeterminacy is when the combined functional role of one sequence of mental states $M_1, M_2, \ldots$ (e.g., a system of beliefs and desires) is functionally indistinguishable from the combined functional role of a distinct sequence of mental states $M_1', M_2', \ldots$ (an alternative system of beliefs and desires), with respect to how the relationships those mental states share with one another and with the rest of the world. This is what gives rise to the problem of functionally isomorphic roles mentioned earlier. In this case, there can in principle be no unique best-fitting scheme of interpretation, regardless of how we choose to measure overall fit and regardless of whether any perfect scheme exists. Any scheme that assigns the states $M_1, M_2, \ldots$ to $P_1, P_2, \ldots$ respectively will fit exactly

1. Or better: if a scheme assigns $P_1, P_2, \ldots$ to $M_1, M_2, \ldots$ 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 mental state M_1 only in the derivative sense that if the $P_1, P_2, \ldots$ jointly occupy the combined functional roles of $M_1, M_2, \ldots$, 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 assign M to P even where P poorly occupies the role of M, if that makes i more fitting overall.

as well as some alternative scheme that assigns $M_1', M_2', \ldots$ to $P_1, P_2, \ldots$ instead, holding all else fixed, because any such a substitution will preserve all the relations that go into deciding fitness. Where this kind of problem arises for a functionalist theory, the only solutions are to either find some further functional roles that might help tell $M_1, M_2, \ldots$ apart from $M_1', M_2', \ldots$ (*à la* Stalnaker), or to impose further restrictions on what kinds of mental states are really possible and thus rule out some of the competing assignments (*à la* Lewis).

But there's still another source of indeterminacy. Assume for the sake of argument that there exists at least one perfect scheme of interpretation—so there's no indeterminacy of compromise, and no functionally isomorphic roles. This implies that if i is perfect, and i assigns $M_1, M_2, \ldots$ to $P_1, P_2, \ldots$ respectively, then there's no way to create an equally fitting scheme by assigning some other mental states $M_1', M_2', \ldots$ to $P_1, P_2, \ldots$ instead. (If we could, then the combined functional roles of $M_1', M_2', \ldots$ would be isomorphic to the combined functional roles of $M_1, M_2, \ldots$, and we've assumed away any such isomorphisms.) However, the assumption does not imply that i is uniquely perfect, for there may yet be an equally fitting scheme i' that assigns $M_1, M_2, \ldots$ to $P_1', P_2', \ldots$ rather than to $P_1, P_2, \ldots$. This will occur whenever the $P_1', P_2', \ldots$ relate to one another and to the world in the same way that the $P_1, P_2, \ldots$ relate to one another and to the world, and this has yet to be ruled out. Roughly: the *problem of functionally isomorphic roles* arises when there are distinct mental states that have functionally isomorphic functional roles, and thus schemes of interpretation can vary in what mental states they identify with a given physical state whilst sharing the same degree of fit overall; by contrast, the *problem of functionally isomorphic occupants* arises when there are distinct physical states that occupy the same functional roles, and thus schemes of interpretation can vary in what physical states get associated with a given mental state whilst sharing the same degree of fit overall.[2]

They are different problems, and a solution to the one need not be a solution to the other. The Lewisean solution—using *naturalness* to constrain what kinds of beliefs and desires are really possible—is a neat way to avoid the former problem, for instance, but it's not going to help with the latter. That's not yet an argument that analytic functionalism in fact faces a problem of functionally isomorphic occupants. We'll get to that in §4. But first, we need to talk about typicality.

3. Anti-Individualism and Typicality

Folk psychology is not about individuals, but about populations. The theory does not tell us anything *directly* about how any particular agent will behave if they're in this-or-that mental state under such-and-such circumstances. Instead, it tells us about how we typically behave under different conditions. We expect that there

2. Readers may recognise that a version of the problem of functionally isomorphic occupants can be seen in Putnam's (1988: 120–5) objection to machine-state functionalism. Putnam's problem arises because there are indefinitely many ways of typing together the physical states with which the machine states of a finite-state machine could be identified, and so any physical system of sufficient complexity can be said to implement any such machine. However, Putnam's version of the problem applies specifically to the nowadays all-but-defunct machine-state functionalism (cf. Chalmers 1996: 323–6). My version applies to anti-individualistic varieties of functionalism, including but not limited to Lewis', which are still very much going strong.

will be some exceptions to the generalisations that folk psychology lays down; that there may be is an important prediction of folk psychology itself. The functional roles associated with the mental states posited by folk psychology should therefore be thought of as kind-relative *typical* roles. This anti-individualism is a consequence of analytic functionalism, not an optional extra.

Suppose that George is a total paralytic. He is in pain, but while his state of being in pain has the right sorts of causes that we'd normally associate with one who's in pain, it has none of the usual behavioural effects. Do not be tempted to say that George is therefore in some state that 'imperfectly occupies' the functional role of pain. That would be misleading. 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). It follows that should any 'imperfect deservers' exist, then folk psychology must be false. But George's paralysis does not *falsify* folk psychology, because folk psychology is consistent with the existence of exceptions. It follows that if we are to properly characterise what it is for a state to *perfectly* occupy the kinds of functional roles defined by folk psychology, then that characterisation needs to be compatible with exceptions. For this we require population-relative typical roles.

Do not be tempted, either, to say that folk psychology is false *for George*, unless this is being used in the loose way to mean merely that George is an exception to the norm. Truth or falsity is not relative to an individual; if a psychological theory is false for George then it's false for everyone. It would be more coherent to say instead that we might have a separate folk-psychological theory indexed to each individual, such that we might say folk-psychology-for-George is false, even while folk-psychology-for-Lennie may yet be true. Coherent, but also mistaken. If our mental states are characterised by the roles they play within a psychological theory, and every individual is associated with a distinct and independent version of folk psychology, then there's no such thing as *pain*—there's just *pain-for-George* and *pain-for-Lennie*. And that is simply not how we conceive of folk psychology, for that theory predicts that George and Lennie can be in the *same* state. To say this, we need a theory that applies to both.

So let's make a distinction. For any type of state S—whether it be a mental state or a physical state—we'll say that S *verifies* the functional role associated with some mental state M (i.e., the M-role) *for an individual* to the extent that, when that individual is in S, then their being in S is caused by and has the sorts of effects as given by the M-role. Verification is always relative to an individual: if George is paralytic then his being in pain does not fully verify the *pain*-role, while Lennie's being in pain might. On the other hand, we want to say that a state *occupies* a role always relative to some reference population, and that George could be in a state that does not verify the *pain*-role for him even while the very same state *perfectly* occupies that role for the kind to which he belongs.[3]

3. If this sounds confusing, then consider: for S to *perfectly* occupy a role *just is* for that state to behave in the manner that folk psychology predicts such states should—and folk psychology predicts that there may be exceptions to its own norms. No more than a few, to be sure, but in a large enough population likely not none. Given this, we should not automatically prefer a scheme of interpretation that assigns P to M only when P fully verifies the M-role for every individual—for that is a scheme that brooks no exceptions even where such exceptions are expected.

Clearly, whether a state occupies a role for a population will have something to do with whether and how well that state verifies the role for the individuals in that population. But much hinges on how we spell this out. Lewis himself never said much on the matter, 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)

The definition implies that S can perfectly occupy a role for a population even while S does not verify that role for each individual. This is the right result. However, there's two problems with Lewis' definition that are worth discussing.

Problem one: the 'with few exceptions' is ambiguous. On one reading, Lewis is saying that S perfectly occupies the M-role for a population just when it fully verifies the M-role for almost every individual—i.e., almost everyone is completely typical, so no more than a few are atypical to any degree. On the other reading, Lewis is saying that S perfectly occupies the M-role for a population just when it at least mostly verifies the M-role for every individual—i.e., perhaps everyone is slightly atypical, but no one is ever highly atypical. An example will help. Suppose that the state of being in pain is associated with many effects: those in pain will say 'ouch', and they will wince, and 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 *most* humans, if they are in pain then they will do *all* of the following: wincing and flinching and saying 'ouch' and ….
2. For *every* human, if they are in pain then they will do *most* of the following: wincing or flinching or saying 'ouch' or ….

Which should we go with? Neither—they're both too strong. The former posits an unrealistic uniformity in the behaviour of mental states across almost every individual in a given kind, whereas the latter implies that the truth of folk psychology is inconsistent with existence of the total paralytic who's in pain. But they also each get something right. It's more plausible that folk psychology tells us that for at least most people, if they're in some mental state then their being in that state will at least mostly verify the typical role associated with that state. To borrow some terminology from Lewis' (1980), we say that a person is *mad* just when, and to the extent that, they deviate from the folk-psychological norm—when their particular way of being in a given mental state M does not verify the M-role. Then if we want a scheme of interpretation that fits as close as possible with what folk psychology has to say, we should be on the lookout for a scheme that's consistent with the hypothesis that we're all at least little mad some of the time, while some few of us might be much more mad than others.

Problem two: Lewis' definition neglects an important distinction between two kinds of 'typicality' hypotheses that exist in folk psychology. 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, we can use facts about an agent's mental states to predict how they're likely to behave. Well call these *predictive hypotheses*. According to the second, we can use about an agent's behaviour to formulate plausible attributions of mental states. We'll call these *interpretive hypotheses*.[4]

Predictive and interpretive hypotheses need not go hand-in-hand. You cannot reliably infer a predictive hypothesis from an interpretive hypothesis, nor vice versa, for the simple reason that 'most As are Bs' does not imply 'most Bs are As'. Imagine, for example, an almost-Super Spartan, who usually avoids any and all expressions of pain but for the occasional lapse upon the stubbing of a toe (cf. Putnam 1980). When an almost-Super Spartan acts as if they're in pain then they are in pain, but when they're in pain they usually do not act as if they are. In the other direction, imagine the perfect actor who always acts as if they're in pain, but is rarely in any actual pain: if they're in pain then they will act as such, but usually aren't in pain when they behave as if they are.

This distinction is important, for it impacts on how we ought to understand *fit*. Let's draw a corresponding distinction between two kinds of roles. 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 at least most of the kinds of causes and effects given by the M-role. Then, supposing that the M-role is unique—i.e., supposing that there are no other mental states M' such that the M-role and the M'-role are functionally indistinguishable—then we can 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 at least most of the kinds of causes and effects given by the M-role, then they are in S.[5] A predictive hypothesis is then a claim to the effect that M occupies the predictive M-role, whereas an interpretive hypothesis is a claim to the effect that M occupies the interpretive M-role.

Given that, Lewis' definition would make sense if folk psychology were comprised solely of *predictive* hypotheses—for in that case, the scheme that fits best with what folk psychology has to say will be the one that, to the maximal extent possible, assigns M to P only if P occupies the predictive M-role. But as Lewis himself was surely aware, we regularly make use folk psychology in both a predictive capacity and in an interpretive capacity. (The theory of belief and desire in Lewis' 1974, for example, rests heavily on the interpretive application of folk psychology.) Better, then, to find a scheme that, to the extent possible, assigns M to P only if P occupies both the interpretive and predictive M-roles.

4. Hiddleston (2011: 400) draws a similar distinction between two ways of parsing 'P typically causes effect e in population p'—between 'it's typical among Ps in p that they give rise to e' (predictive) and 'it's typical among the causes of e in p that they are Ps' (interpretive).

5. Where the M-role isn't unique, matters are more complicated. Suppose that M and M' have functionally isomorphic roles. Then we can say that $S, S', \ldots$ *jointly* occupy the *combined* interpretive M-and-M'-and-…-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 most of 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' or …. If we continue to assume that the problem of functionally isomorphic roles has already been solved somehow, then we don't have to worry about this more complicated formulation.

4. Functionally Indistinguishable Occupants

The best scheme of interpretation, if there is just one, will be the one that best fits with what folk psychology has to say. The problem is that if folk psychology is comprised of statistical typicality hypotheses like those we've just been discussing, then there won't be just one best scheme of interpretation.

Consider first a fictional example. Imagine a small population of just six mice. According to our 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 behaviour respectively. After a bit of (humane) experimentation, we come to find that mouse brains are incredibly simple: they each have just five neurons, with each neuron belonging to one of two clearly distinct types that we label the a-type and the b-type. Each has a different proportion of these than the others. Furthermore, we find that whether a mouse behaves in an a-like way or an b-like way corresponds to whether most of their neurons are a-type or b-type. Consequently, there are six *specific* states we distinguish, one for each mouse, that we label P_1 through P_6:

$$\begin{array}{lcll} P_1 & = & a,\ a,\ a,\ a,\ a & \Rightarrow \quad a\text{-like behaviour} \\ P_2 & = & a,\ a,\ a,\ a,\ b & \Rightarrow \quad a\text{-like behaviour} \\ P_3 & = & a,\ a,\ a,\ b,\ b & \Rightarrow \quad a\text{-like behaviour} \\ P_4 & = & a,\ a,\ b,\ b,\ b & \Rightarrow \quad b\text{-like behaviour} \\ P_5 & = & a,\ b,\ b,\ b,\ b & \Rightarrow \quad b\text{-like behaviour} \\ P_6 & = & b,\ b,\ b,\ b,\ b & \Rightarrow \quad b\text{-like behaviour} \end{array}$$

What we'd like to do is partition $\{P_1, \ldots, P_6\}$ into two more general types—we'll label them P_a and P_b—in such a way that there exists a fitting scheme of interpretation i such that $i(P_a) = M_a$ and $i(P_b) = M_b$.

There's one obvious way to do this. The states P_1–P_3 jointly correspond to the state of *having mostly a-type neurons*, which occupies the predictive and interpretive M_a-roles. The states P_4–P_6 jointly correspond to the state of *having mostly b-type neurons*, which occupies the predictive and interpretive M_b-roles. So where '$P_x = P_y \vee P_z$' means that a mouse is in P_x if and only if it's in either of the more specific states P_y or P_z, the most straightforward solution is to have:

TYPING 1. $P_a = P_1 \vee P_2 \vee P_3, \quad P_b = P_4 \vee P_5 \vee P_6$

But there's other ways of dividing up P_1–P_6 that would also work:

TYPING 2. $P_a = P_1 \vee P_2 \vee P_6, \quad P_b = P_4 \vee P_5 \vee P_3$

TYPING 3. $P_a = P_1 \vee P_5 \vee P_3, \quad P_b = P_4 \vee P_2 \vee P_6$

TYPING 4. $P_a = P_4 \vee P_2 \vee P_3, \quad P_b = P_1 \vee P_5 \vee P_6$

Assuming that 'few exceptions' here means no more than one exception, then each of the following holds for all four of the above:

PREDICTIVE. With few exceptions, if a mouse is in P_a (or P_b), then it will behave in an a-like way (or b-like way, respectively).

INTERPRETIVE. With few exceptions, if a mouse behaves in an a-like way (or b-like way), then it is in P_a (or P_b, respectively).

But now we face indeterminacy: for each individual mouse, there exists at least one fitting scheme according to which that mouse is in M_a, and another equally-fitting scheme according to which that it's in M_b.[6]

The basis of the problem is fit depends on typicality facts, and typicality facts in turn depend on what's true of a population in a way that's usually robust against minor variations in the exact makeup of that population—it doesn't so much matter *who* the exceptions are, just that there are *few* of them. Think of each distinct way of dividing up the specific P_1 through P_6 into the more general types P_a and P_b as a potential way of precisely specifying the sub-populations of mice that are taken to be in either M_a or in M_b. So, for example, if we go with Typing 1 then the sub-population of mice in M_a will be those three in P_1, P_2, or P_3, whereas if we go with Typing 2 then it consists of those three in P_1, P_2, or P_6. Now provided that these are *mostly* similar to one another, then as far as the specified facts about typicality are concerned it makes no difference whether we opt for one or the other—but it does make a big difference to the *individuals* whether they're counted as belonging to one sub-population or the other.

Consider next an analogy: Lennie lives in New York, while George lives in London. Since both of those cities have a very large population, neither Lennie nor George by themselves make any appreciable difference to what's typical of New Yorkers and of Londoners respectively. Consequently, if they were to swap places suddenly, holding all else fixed, then their doing so would likely make no appreciable difference to the typicality facts either. As far as the statistics are concerned, it doesn't much matter which of the two is a resident of what city—Lennie and George can be freely swapped out for one another, provided that everyone else's location is held fixed. In that specific sense, then, Lennie and George are interchangeable *salva typicalitae*.[7] And what's true of Lennie and George is true of any other pair of residents from New York and London, or at least the very majority of them, for the same reason. If, instead of swapping Lennie and George, we were to swap Candy from New York for Curley from London, it would again make no appreciable difference to the typicality facts. However, it clearly matters a big deal to the *individuals* where they happen to live!

So now the general argument. Suppose we have two mutually exclusive 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 assigns M to P. Lennie is in P, while George is in $\neg P$, and so according to i Lennie is in M and George isn't. 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 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*. Given that, let P' and $\neg P'$ be some alternative pair of mutually exclusive physical states that are almost exactly identical to P and $\neg P$ respectively, except that:

6. Typing 1 through Typing 4 are not exhaustive. There are 12 more ways of defining P_a and P_b that will have them satisfying both the predictive and interpretive roles, and another 6 on which they satisfy just the predictive role—try to find them all!

7. A colleague much more adept at Latin than I am tells me that it should be 'salva tipico'. My apologies for the liberties taken with the language.

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

Provided that the population of individuals in P and in $\neg P$ is large enough, what's typical of those in $P/\neg P$ is likely to be the same as what's typical of those in $P'/\neg P'$. In particular, if P occupies the interpretive and predictive M-roles, then very likely so too will P'. So there will be some other fitting scheme, i', that assigns M to P' instead. But now it's indeterminate for Lennie and George whether each is or is not in M. Now run the same argument for every mental state, and any pair of individuals such that one is in the state and the other isn't, and widespread radical indeterminacy is the result.

5. How Many is Humankind?

The key premise of the foregoing argument is that for any candidate physical state P that can be said to occupy the functional role for some mental state M, there will be so many in P and so many not in P that swapping just one individual from one of these sub-populations to the other will likely make no difference to the relevant typicality facts. And that premise is going to hard to deny on the Lewis' theory, for on that theory the population that constitutes *humankind* includes 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; see also Lewis 1980: 219–20, 1983: 120).

Lewis never explained why he thought the population relative to which the typicality facts for humankind ought to include also the inhabitants of nearby possible worlds, but there are at least three reasons we can note here: (i) we want our schemes of interpretation to be insensitive to irrelevant contingencies, (ii) we want them to be non-trivially constrained for actually uninstantiated 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 have more or less the same scheme of interpretation as our own—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 that Lennie, George, Crooks and Slim are all in some state M, and they are 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 and lives in some location far away from the others. 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. 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 at some faraway location had been cloned a few times—i.e., if nothing at all about Lennie's physical constitution or local environment were altered.

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. Of the absurdly many beliefs and desires we might have, for instance, we can surely have had only a relatively small 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 be 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 or will ever be in with mental states M that no one has ever actually been or will be in, such that, with few exceptions, if we were in the former then we'd be in a state that verifies the roles of the latter. But that is (on the Lewisean view) 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 lone madman. 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 (say), but where this individual's way of being in pain does not verify the *pain*-role particularly well—perhaps they are an extreme masochist and seek out intense pains, which they get upon the tickling of their feet. If the relevant typicality facts are always defined relative to the population of a single world, then this ought to be impossible. But the lone madman is only ever alone relative to their own world. If the reference population can extend out into relevantly similar worlds, then we open up the conceptual 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 similar constitution at nearby possible worlds, is in their specific case behaving in an unusual way.

I'm uncertain 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 did not provide his reasons for saying this, but I think analytic functionalists should disagree since it rules out the possibility of the lone madman. Presumably, he thought that facts about actual populations should have special priority—the core idea of functionalism is, after all, that we find the scheme of interpretation that works best given the facts *of our world*. But this does not mean that we to prioritise the typicality facts of the actual world *considered in isolation*, for we might instead prioritise the typicality facts of a space of nearby possible worlds *centred on the actual world*. What counts as 'nearby' will always depend on what world is actual, so the facts about what's typical relative to the 'nearby' worlds just are facts about the actual world.

In any case, even if only one or two of these reasons are convincing the upshot is the same: it's important for the Lewisean view that the population relative to which the typicality facts are defined is large—*very* large. For humankind, it'll include not just the already enormous population of all humans who have lived and ever will live, but many of counterparts at the many worlds relevantly similar to our own. And relative to such a large population, it's likely that actual individuals in apparently distinct mental states will be interchangeable *salva typicalitae*.

6. Possible Solutions

I'll close by considering three (non-exclusive) ways of responding to the problem. The first two involve minimising madness and maximising naturalness respectively. The third involves reconceptualising the kind of typicality that's most relevant to folk psychology.

6.1 Minimise Madness

Consider again the mouse example. There's something *right* about Typing 1. Perhaps it's the fact that it minimises exceptions: a mouse is in P_a just in case they behave in a mostly a-like way, and a mouse is in P_b just in case they behave in a mostly b-like way. The others fit with the same predictive and interpretive hypotheses, but they do so only by positing rare exceptions where Typing 1 posits none. This suggests a possible solution—a *minimise madness* constraint. Something to the following effect would work: a scheme of interpretation fits better, all else equal, to the extent that whenever it assigns M to P, then P verifies the M-role for the maximal number of individuals in the population and/or maximises average degree to which P verifies the M-role for those individuals.

Lewis *seems* to have suggested something along these lines at one point. 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)

And then a little further on:

> The best scheme is the one that does the best job overall of conforming to the constraining principles, *taking one individual and time with another*. (1983: 120, emphasis added)

But this is very hard to interpret. If the "best interpretation of Karl considered in isolation" does not assign the stated beliefs and desires to him, then there will be a scheme of interpretation for Karl's population that does not assign those beliefs and desires to him and fits at least as well as any scheme that does. Because of this, the *minimise madness* constraint ends up collapsing the distinction between individualistic and anti-individualistic functionalism. For suppose that i is some

fitting scheme that assigns the state of *desiring fame*, say, to some state P that Karl happens to be in. Now 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 predictive and/or interpretive roles for *desiring fame* relative to Karl's population, then P' will occupy those same roles at least as well as P does. But then the *minimise madness* condition will kick in and tell us that (all else equal) we ought to prefer some alternative scheme i' that assigns the state of *desiring fame* to P' instead.

The upshot is that if we opt for a *minimise madness* constraint, then it *cannot* be the case that 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." The correct interpretation for Karl's population will just be the summation the best interpretation of each individual of that population considered in isolation. And that's unacceptable. Anti-individualism does theoretical work for Lewis, and should not be given up easily. The goal is to associate the mental with the physical in such a way as to maximise folk psychology's closeness to the truth, and folk psychology seems to tell us that 'madness' is possible. It tells us that Karl could in principle believe himself a fool and desire fame even while his own evidence and behaviour suggest no such interpretation, because Karl may just behave in a 'mad' way relative to his beliefs and desires. It tells us also that there might be some for whom pain is quite unusual:

> 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)

Minimise madness rules these kinds of cases out; it might work as a solution to the problem of functionally isomorphic occupants, but only at the expense of undermining the very position it's supposed to save.

6.2 Maximise Naturalness

Maybe what's right about Typing 1 is something else. Another salient property of Typing 1 is that it partitions $\{P_1, \ldots, P_6\}$ 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 is for it to have mostly *a*-type neurons, and for a mouse to be in P_b is for it to have mostly *b*-type neurons. By comparison, the other partitions sort P_1–P_6 into types that are intuitively 'disjunctive'. This suggests an alternative solution—a *maximise naturalness* constraint that tells us to prefer a scheme that assigns mental states to more natural physical states whenever possible.

Such a constraint would be a significant addition to the Lewisean theory. It would posit a further role for naturalness that's not implied by either of the two roles that Lewis thought naturalness might play in his theory 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 propositional attitudes:

> It is here that we need natural properties… a bias towards believing that things are green rather than grue, towards having a basic desire for long life rather than for long-life-unless-one-was-born-on-Monday-and-in-that-case-life-for-an-even-number-of-weeks. In short, [naturalness constraints] impute ineligible content, where ineligibility consists in severe unnaturalness of the properties the subject supposedly believes or desires or intends himself to have. (Lewis 1983: 375)
>
> And it [i.e., folk psychology] sets presumptive limits on what our contents of belief and desire can be. Self-ascribed properties [i.e., the contents of our *de se* attitudes] 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 noted above, naturalness is used in this role to solve a specific instance of the problem of functionally isomorphic roles that arises for beliefs and desires, essentially by rendering certain states of belief and desire impossible. (See Lewis 1983: 373–7; 1986: 107–8; see also Schwarz 2014, Weatherson 2012 and Williams 2016; 2018 for discussion.) That problem is independent of the problem raised in §4—trim the possible states of belief and desire down however you like, and you'll still need to contend with the problem of functionally isomorphic occupants for those that remain.

The second (and relatively minor) role for naturalness in the Lewisean theory is as a constraint on admissible grammars, the recursive rules by which indefinitely extended meaningful sentences can be constructed out of finitely many subsentential expressions. (See Lewis 1992; see also Weatherson 2012). We should want grammatical rules that extend out to infinity in a plus-like way, rather than in a quus-like way. It should go without saying that this has nothing to do with the kind of *maximise naturalness* constraint being considered here.

(In 'Putnam's Paradox', Lewis (1984) also described a position that's since come to be known as *reference magnetism*—that we should assign extensions to terms so as to optimise some balance of naturalness versus fit with use. Reference magnetism *does* seem to imply the kind of *maximise naturalness* constraint under consideration, since it will tell us that if two physical states, P and P' both fit equally well with how we use the word 'pain', then 'pain' will refer to the more natural of the two. But Lewis was no reference magnetist. He was quite explicit on more than one occasion that the position outlined in 'Putnam's Paradox' is not his own. (See Lewis 1984: 222 and 1983: 373.) On his *actual* view, the meanings of complete sentences spoken within a linguistic community are determined by the conventions at play within that community, which are fixed in turn by the beliefs and desires of the individuals therein; and with sentential meanings fixed, the meanings of subsentential expressions are then determined by the somewhat arbitrary choice of a grammar that will generate the right kinds of sentential contents. (See especially Lewis 1974, 1975, and Schwarz 2014.) Naturalness only plays a very indirect role in connection to fixing the meanings of our terms, by constraining the kinds of beliefs and desires that are possible and hence the kinds of conventions that might hold in a linguistic community, and by constraining the kinds of grammars we might choose from.)

None of this is to say that we shouldn't impose a *maximise naturalness* constraint on our schemes of interpretation; just that in doing so we'd be adding something to the Lewisean theory. Maybe it's the right solution. Or maybe a better solution still would be to combine *minimise madness* and *maximise naturalness*, so that schemes of interpretation optimise some balance of minimising exceptions and maximising the naturalness of the physical states with which our mental states end up being identified. I'm not going to try to convince you that none of these ideas would work, but I do want to say that there's yet one more potential solution that is altogether more plausible. Perhaps the real problem is that we've been thinking about folk psychology in the wrong kind of way.[8]

6.3 Rethinking Typicality

Consider Lewis' defective calculator example:

> We have a certain kind of mass-produced calculator: the Texas Instruments 58C, let us say. A certain hardware state S of the 58C is the state of having the number 6099241494 stored in register 17. The state is to be thus interpreted because of its causal role in the functional organisation of the 58C. Now suppose that one 58C comes out defective. It still *is* a 58C… but the defect does mean that the role of S in this calculator differs from the normal role of S in the 58C. Considering the defective calculator in isolation, there is no reason to interpret S as the state of having 6099241494 in register 17. But that is how we interpret S for the 58C generally, so that is what S is for any 58C—even the defective one. (1983: 119; see also 1986: 39–40)

Lewis suggests the reason why we're inclined to interpret S as the state of having 6099241494 in register 17 is that that's the statistically normal causal role of S. But there's another explanation: we're inclined to interpret S as the state of having 6099241494 in register 17 because that's its proper function, given how the calculator was designed. The defective calculator is defective not because it's statistically abnormal, but because it's not working the way it's supposed to work. Imagine if the defect were the first one ever produced, and production were shut down before any more could be made. Still we would be inclined to interpret the lone defect's S as the state of having 6099241494 in register 17.

8. There's at least two concerns for the *maximise naturalness* solution. I'll mention them here in a footnote to emphasise that ruling out the *maximise naturalness* solution is not the goal of the present discussion. First problem: the constraint is only meaningful if we have an account of what it is for one physical state of an agent to be more or less natural than another. I not know what such an account would look like; I do not envy the person who tries to develop one. Second problem: the constraint may not play nicely with the fact of neuroplasticity. Consider, for example, 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 characterise 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.

As with the calculator, so too with us. Lewis mixed up two kinds of normativity. To describe something as 'typical' is to say that it exemplifies the distinctive qualities of some normative type. So far in this essay, the relevant type relates to statistical norms: something counts as typical just when it conforms well to expectations given what others in the population to which it belongs are like. But our folk theories do not trade only in, or even primarily in, statistical norms. There is also the non-statistical kind of 'typicality'—we'll call it *archetypicality*—that relates to what members of a kind are ideally supposed to be like when everything is functioning in the way it's supposed to function. That, I claim, is the kind of typicality that we should be focusing our attention on.

Let me spell that out a bit more. As Griffiths (2002) has argued, an extremely widespread and cross-cultural feature of pretheoretic biological thought is *folk essentialism*—the idea that all living things manifest, and strive to manifest, an underlying essence (or an *archetype*) that characterises the kinds to which they belong. This archetype is normative in the sense that it defines what members of the kind ideally ought to be like, and how their various parts are supposed to function, whereas real instances of the kind will be at best imperfect 'copies' or 'reflections' of the ideal form. Moreover, this archetypal norm is logically independent of any statistical norms. It is presumed that real instances of a given kind will innately *strive* towards being more like their respective archetypes; the proper process of human development, for example, is always directed towards being more like the archetypal human, though sometimes the world gets in the way and diverts us from this course. Statistical norms thus fall out of folk essentialism *given* certain auxiliary hypotheses: it implies that we can expect most members of a kind will approximate their archetypes to a good degree *if all goes well*. But all need not go well. Were some terrible pandemic to cause all humans to be born without the sense of taste, such that being tasteless were now the statistical norm, still the archetypal human would have a working sense of taste.

What's true of folk biology is likely true of folk psychology. After all, if human physiology is understood by reference to some ideal human archetype, then why not also human psychology? Consider that part of folk psychology that deals with beliefs and desires in particular, where rational norms abound. The thesis of folk psychology is not in the first instance that almost all humans do in fact have rational beliefs and desires given the evidence of their senses, and that they will almost always behave in such a manner as would maximise the satisfaction of their desires were all their beliefs true—it is rather the thesis that this is what our beliefs and desires *should* be like, how they *should* arise in response to evidence and how they *should* interact with one another to produce intelligent behaviour. Folk psychology no doubt also predicts that most of us will probably approximate those rational norms to a greater or lesser degree, assuming that nothing drastic has interfered with the proper course of our development and day-to-day functioning. But this is not so clearly a part of what *defines* our states of belief and desire—we can readily imagine cases such as the lone madman, or even whole populations of mad people, just as we can imagine an entire population of defective calculators. The role our beliefs and desires play is not in the first instance causal-statistical, but causal-archetypal.

The argument of §4 presupposes that a scheme of interpretations degree of fit is to be characterised wholly in statistical terms. The problem arises because we can swap and change around what individuals are taken to be in which mental states without affecting the population-level statistical facts, provided that the population relative to which those facts are defined is large enough. But if folk psychology is primarily in the business of positing archetypal norms, then the argument of §4 no longer works. The size of the population and what the individuals within that population are in fact like is irrelevant, since the goal is to not to ensure that M is identified with P only when P verifies the M-role for *most* agents in the population. Rather, we should instead be on the look out for schemes of interpretation that assign mental states M to physical states P such that the way the latter are *supposed* to behave aligns as neatly as possible with the way the former are *supposed* to behave according to folk psychology.

Of course, if this is going to work then we will need to be able to talk about physical states having some proper function or design—something they're *supposed* to do—that's independent of any hypothetical identification of those states with mental states. I take it that this problem has been more or less solved in the biological case. Indeed, it's been solved twice over; once by the aeteological analysis of 'proper function' and 'design' that characterises the function of a trait in terms of the effects for which it has been naturally selected (Millikan 1989, Neander 1991), and once by the propensity analysis that characterises the function of a trait in terms of the effects which render the trait adaptive (Bigelow & Pargetter 1987). Both provide us with perfectly respectable accounts of 'function' that can be applied to wholly physical systems, and can be used to make sense of one neurophysiological state having a proper function.

(This is not to say that folk psychology *itself* characterises what our mental states are supposed to be like in terms of effects selected for through cumulative natural selection or biological fitness-enhancement. As Adam Pautz notes, such things 'would be far too *recherché* to be something that the folk implicitly know a priori' (2013: 223). He takes this to mean that the analytic functionalist cannot make appeal to the more sophisticated resources of modern biological theory when characterising the functional roles our mental states are supposed to play within folk psychology. And that's correct—but the naturalised kind of functions we find in those modern biological theories can still serve as the *best deservers* of the kinds of normative causal-archetypal relations posited by our folk theories.)

The suggestion, then, is that the best scheme of interpretation—if there is just one—will be the one that assigns M to P such that, to the degree possible, the (naturalistic) function of P matches up neatly to the archetypal role that M is supposed to play according to folk psychology. There might be more than one best scheme. Likely there will be some indeterminacy still how we precisely specify the physical states, such that this-or-that state can be said to have been naturally selected for this-or-that purpose, or to have such-and-such adaptive effects. But indeterminacy is not inherently problematic, and I see no reason to suppose that a bit of leeway in which physical states get to count as having what proper biological functions will lead to any *radical* indeterminacy. At the very least, on this picture the argument of §4 gives us no cause for concern.

References

Bigelow, John, and Robert Pargetter. 1987. "Functions." *Journal of Philosophy* 84 (4): 181–196.

Chalmers, David. 1996. "Does a rock implement every finite-state automaton?" *Synthese* 108:309–333.

Griffiths, Paul E. 2002. "What Is Innateness?" *The Monist* 85 (1): 70–85.

Hiddleston, Eric. 2011. "Second-order properties and three varieties of functionalism." *Philosophical Studies* 153:397–415.

Lewis, David. 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 Beebee and A.R.J. Fisher. Oxford: Oxford University Press.

Millikan, Ruth Garret. 1989. "In Defense of Proper Functions." *Philosophy of Science* 56 (2): 288–302.

Neander, Karen. 1991. "Functions as Selected Effects: The Conceptual Analyst's Defense." *Philosophy of Science* 58 (2): 168–184.

Pautz, Adam. 2013. "Does Phenomenology Ground Mental Content?" In *Phenomenal Intentionality,* edited by Uriah Kriegel, 194–234. Oxford.

Putnam, Hilary. 1980. "Brains and Behavior." In *Readings in Philosophy of Psychology,* edited by Ned Block, 1:24–36. Harvard University Press.

———. 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.

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.

———. 2019. *The Metaphysics of Representation.* New York: Oxford University Press.