Many philosophical naturalists eschew analysis in favor of discovering metaphysical truths from the a posteriori, contending that analysis does not lead to philosophical insight. A countercurrent to this approach seeks to reconcile a certain account of conceptual analysis with philosophical naturalism; prominent and influential proponents of this methodology include the late David Lewis, Frank Jackson, Michael Smith, Philip Pettit, and David Armstrong. Naturalistic analysis (sometimes known as "the Canberra Plan" because many of its proponents have been associated with Australian National University in Canberra) is a tool for locating in the scientifically given world objects and properties we quantify over in everyday discourse. This collection gathers work from a range of prominent philosophers who are working within this tradition, offering important new work as well as critical evaluations of the methodology. Its centerpiece is an important posthumous paper by David Lewis, "Ramseyan Humility," published here for the first time. The contributors first address issues of philosophy of mind, semantics, and the new methodology’s a priori character; then turn to matters of metaphysics; and finally consider problems regarding normativity. "This volume shows how the so-called Canberra Plan of metaphysical research continues to inspire (and provoke) some of the most interesting work in modern metaphysics. The collective range of its contents, all published here for the first time, is as impressive as the depth, rigor, and clarity of the individual chapters. The authors’ contributions and the editors’ lucid introduction constitute an indispensable conspectus of the plan’s progress and present prospects."

—Hugh Mellor, Emeritus Professor of Philosophy, University of Cambridge

"In this excellent volume, David Braddon-Mitchell and Robert Nola have brought together new essays by the most important contributors to the debate about the viability of the Canberra Plan, both defenders and skeptics. Anyone with a serious interest in metaphysics, a priori knowledge, the philosophy of mind, and metaethics will need to possess this book."

—Tim Crane, Professor of Philosophy, University College London

"Though the method of conceptual analysis isn’t much in favor these days, the essays in this volume show just how far our understanding of the method has come, how misguided many of the criticisms of the method have been, and how productive the method can still be. The essays, which include David Lewis’s already much discussed though not previously published ‘Ramseyan Humility,’ make compulsory reading for everyone with a serious interest in how philosophy at its most exacting is done."

—Michael Smith, Department of Philosophy, Princeton University

David Braddon-Mitchell is Associate Professor of Philosophy at the University of Sydney and the author (with Frank Jackson) of The Philosophy of Mind and Cognition. Robert Nola is Professor of Philosophy at the University of Auckland and the author of Rescuing Reason.
Conceptual Analysis and Philosophical Naturalism
Dedicated to the memory of David K. Lewis
Contents

Acknowledgments ix

1 Introducing the Canberra Plan 1
David Braddon-Mitchell and Robert Nola

I Mind, Concepts, and Theories 21

2 Naturalistic Analysis and the A Priori 23
David Braddon-Mitchell

3 Folk Psychology and Tacit Theories: A Correspondence between
Frank Jackson, and Steve Stich and Kelby Mason 45
Frank Jackson, Kelby Mason, and Steve Stich

4 A Priori Biconditionals and Metaphysics 99
Frank Jackson

5 The Argument from Revelation 113
Daniel Stoljar

6 Names, Plans, and Descriptions 139
Fred Kroon

7 Jackson’s Armchair: The Only Chair in Town? 159
Justine Kingsbury and Jonathan McKeown-Green

8 Is Semantics in the Plan? 183
Peter Menzies and Huw Price

II Metaphysics 201

9 Ramseyan Humility 203
David Lewis
10  A Partial Defense of Ramseyan Humility  223  
Dustin Locke

11  Physicalism without Pop-out  243  
Philip Pettit

12  Platitudes and Metaphysics  267  
Daniel Nolan

III  Normativity  301

13  Naturalizing Normativity  303  
Mark Colyvan

14  Moral Functionalism, Ethical Quasi-Relativism, and the Canberra Plan  315  
Denis Robinson

References  349
Contributors  363
Index  365
Acknowledgments

We would like to acknowledge the permission to publish David Lewis’s paper “Ramseyan Humility” given by the executors of the Estate of David Kellogg Lewis. In particular we would like to thank Steffi Lewis for her generous assistance in making possible the inclusion of the paper in the collection.

The philosophical program discussed in this book was largely inspired by the work of David Lewis. We would like to dedicate this book to his memory.

None of the papers published in this volume has appeared elsewhere before.
1 Introducing the Canberra Plan

David Braddon-Mitchell and Robert Nola

1 Introduction

This collection of essays is devoted to a critical evaluation of a project of philosophical methodology and analysis known colloquially as the ‘Canberra Plan’. The original driving forces behind the project were David Lewis and Frank Jackson (see Lewis 1970, 1972; Jackson 1994b, 1998a) and some of the many people who were associated with the Philosophy Program of the Research School of Social Sciences at the Australian National University in Canberra during the 1990s. The first published use of the expression ‘Canberra Plan’, however, was by two of its critics (O’Leary-Hawthorne and Price 1996, 291). Since Canberra is a planned city founded originally as the seat of the Federal Government of Australia, its detractors complain that it lacks the features that arise in cities that grow organically and have diverse inhabitants who are not largely government bureaucrats. Originally an ironic suggestion, the metaphor turns on the allegation that the Canberra Plan adopts a view of language that misses its functional diversity. As the essays in the collection show, the label has been “rescued” and adopted by those who endorse the Canberra Plan.1 However, it should be emphasized that not all “Canberra Planners” agree on all aspects of what the project should be, as the essays in this volume reveal.

The project has its origin in Frank Ramsey’s paper “Theories” (Ramsey 1990, 112–136). Ramsey gives an example of a theory, but it is not the standard kind of theory one finds in the sciences. Rather, it is a somewhat arcane theory intended to capture what it is like to move backward or forward with one’s eyes open or shut while noting what one’s perceptual experiences are (of nothing, of color, etc.). The details of this theory need not detain us, but two points arise from it. The first is that a theory can be about any subject matter, including that of Ramsey’s odd example. This
is a point not lost on Canberra Planners who apply their project over many domains including not only the sciences broadly understood to include the physical, psychological, and social, but other domains such as “folk” systems of belief, or the normative (including the rational and the moral), and so on. The diversity of applications is reflected in the essays in this volume.

The second is this. Ramsey distinguishes between two kinds of terms in his theory. Despite the fact that his example is about perceptual items, his distinction is not the old epistemological distinction advocated by positivists between terms denoting observables and terms denoting nonobservables (or terms denoting sense data and terms denoting the nonsensory, etc.). He leaves quite open how the distinction is to be drawn and simply refers to the two kinds of vocabulary as that of a primary system and secondary system. This again is a point not lost on Canberra Planners; as will be seen, they adopt a wide variety of ways of making the distinction between the two kinds of vocabulary, very few of which are redolent of the old positivistic distinction. They follow a suggestion made by David Lewis in his “How to Define Theoretical Terms” (1970; see also Lewis 1999 [1972], 250). The vocabulary of a theory is divided into T-terms and O-terms. The O-terms are not necessarily observations terms; they could be other terms (other than those distinctive of the theory), or old terms (old because we are already acquainted with their meaning). The important difference is that whereas the O-terms get their meaning, in whatever way, from outside the theory, the T-terms get their meaning from the role they play within the theory and are implicitly defined in the context of the theory. For that reason we can also call the O-terms outsider terms while the T-terms are insiders, defined implicitly by the theoretical context in which they occur. In what follows we will adopt the labels ‘O-terms’ and ‘T-terms’, taking this to be closely akin Ramsey’s distinction between, respectively, the vocabularies of the primary and secondary systems.

Ramsey is also famous for a suggestion that occupies only a few pages of his paper but which has acquired considerable importance. His suggestion is that a theory can, in a certain sense, be replaced by another expression in which each of the O-terms is retained but the T-terms are replaced by variables, with an existential quantifier binding each of the variables. Hempel seems to have been the first to call the new sentence the “Ramsey sentence” in his 1958 paper “The Theoretician’s Dilemma” (Hempel 1965, 216). Though not logically equivalent, the original theory and its Ramsey sentence version are equivalent with respect to what can be said in the
vocabulary of O-terms (including, of course, logical vocabulary, something we will assume from now on without explicitly saying so).

To illustrate, if ‘electron’ is a T-term in some theory, it can be replaced by a variable, \( x \), bound by an existential quantifier (and so on for all other T-terms in the theory such as ‘charge’, ‘spin’, etc., which can be replaced by variables \( y, z \), etc.). In this sense the vocabulary of T-terms is eliminated while the vocabulary of O-terms is retained. But no entities are thereby eliminated. There is an intended domain over which the variables range. The Ramsey sentence says that there is “something,” \( x \), in the domain of the world that satisfies the first of two conditions: (a) all (or most) of what is said of \( x \) is couched in the vocabulary of O-terms. A story is told about the role of \( x \) using just O-vocabulary.

If there is not just one T-term but several, such as ‘charge’ and ‘spin’, which are replaced by existentially bound variables \( y \) and \( z \) respectively, then a further condition obtains: (b) the “something” \( x \) in the domain of the world also satisfies the “structural” relations that the theory postulates as obtaining between it and other “somethings,” such as \( y \) and \( z \), in the domain of the world. This is a point emphasized by Carnap (1966, chapter 26, 252) even though he is not an advocate of the realism that such a remark might be taken to entail. In contrast, if the term ‘phlogiston’ occurs in a theory then we can be eliminativists about phlogiston. The Ramsey sentence of this theory, which is obtained by replacing the term ‘phlogiston’ in that theory by a variable, is false: there is no “something,” \( x \), in the world that satisfies all (or most) of what can be said of it in the vocabulary of O-terms and stands in any requisite “structural” relations to any other “somethings” (\( y, z \), etc.).

To take an even simpler example, we might end up with an expression that says something like a conjunction of the following (or at least would, given a crude and false theory of the atom). There is one kind of thing, and another, and yet another; instances of the first of these orbit a clump of instances of the other two; instances of the first and the second are attracted to each other; instances of the first repel each other, as do instances of the second; instances of the third exhibit no attraction or repulsion to other instances of its own kind; some strange force keeps the members of the second kind together in a clump despite their mutual repulsion; and so on. Such an expression introduces no terms other than those of an earlier and familiar scientific vocabulary of preatomic science, yet if it were true, there would be classical atoms. The reader can readily guess that this story is about electrons, protons, and neutrons. But the point about how the story has just been told is that we do not need to
employ any new T-terms such as ‘electron’, ‘proton’ and ‘neutron’ at all. Everything the story describes can be done without them by employing only the old, well-known vocabulary of O-terms that we understand antecedently, along with existential quantifiers. That is, we can equally as well get by with the Ramsey sentence version of the theory as we can with the theory itself, as far as the common O-vocabulary content of both is concerned. We do not need the theoretical terms, as we can do just as well in telling the scientific story using the existentially quantified bound variables in their place.

After Ramsey’s initial paper there were two significant developments of his approach to theories. The first is due to Carnap; but we can set this aside as not germane to the development of the Canberra Plan. The second development is due to several papers by David Lewis, especially “How to Define Theoretical Terms” (Lewis 1970); we focus on this here as it provided the initial impetus for the Canberra Plan. Lewis accepts much of the framework that Ramsey suggested but he made some significant additions. One seemingly minor change is the following. The existentially quantified Ramsey sentence says that there is at least one “something” in the world satisfying appropriate clauses along the lines of (a) and (b) above. Assuming the Ramsey sentence is true, then in the case of just one existentially bound variable in the Ramsey sentence the “something” that does the satisfying is a single entity, say, A. In the case of two existentially bound variables the “somethings” that do the satisfying are listed as an ordered pair of entities, ⟨A, B⟩; and so on, from ordered triples up to the n-tuple ⟨A, B, . . . , N⟩ (in the case of n existentially bound variables).

But the existentially quantified Ramsey sentence leaves open the possibility that there might be more than one entity as satisfier, or pair of entities (or triple, etc.) of satisfiers. The possibility of multiple realization (often an alternative expression for ‘satisfaction’) is an issue that Lewis addresses explicitly in his essay in this volume. But in his 1970 paper this possibility is closed off since he argues there that our best scientific theories will be uniquely realized. That is, it is required that there be no more than one entity (or pair, or triple, etc.) realizing the requisite conditions along the lines of (a) and (b). In effect, the existential operator is replaced by a definite description operator, yielding a generalized definite description (or a number of them, depending on the number of variables bound by existential operators in the Ramsey sentence). On the standard view of definite descriptions, if a unique “something” is picked out then the description denotes that “something”; but if there is no “something,” or more than one “something,” then the description fails to denote. (Later Lewis modi-
fied his view about what to say when there are two or more equally good denotata.6) The details of his earlier view are set out in “How to Define Theoretical Terms” (Lewis 1970) and “Psychophysical and Theoretical Identifications” (Lewis 1972), and they are mentioned in various essays in this volume (particularly chapters 9, 10, and 11). In this respect Lewis provides an answer to a matter raised in Ramsey’s “Theories,” namely, how to give an “explicit definition” of each T-term; this is done using a generalized definite description in which only O-terms occur (and, as implicitly understood, the vocabulary of logic).

Given the apparatus provided by Ramsey and Lewis, we can now give a further simple illustration (suggested by Lewis 1972) of how it may be used. Consider how detectives might build up a story, or theory, based on the evidence they collect concerning, say, the 7 July 2005 bombings in London that occurred in three underground trains and a bus. Initially, massive amounts of information from eyewitnesses, Closed Circuit TV, cameras in mobile phones, and the like become available, which can be described in the vocabulary of already well-defined “outsider” terms, the O-terms of our ordinary language. A story emerges in which a person, for whom we can introduce the name ‘X’, causes, in some manner, such-and-such happenings in one underground train (here the story is couched in O-vocabulary terms of ordinary language); another story emerges in which some other person, whom we can name ‘Y’, causes, in some manner, so-and-so happenings on a bus (here there is a further story also couched in O-vocabulary terms); and so on for the other persons named ‘Z’ and ‘W’. Eventually the information evolves into a story about all four conspirators X, Y, Z, and W, who we can say are the “role players” in the story. We know nothing more about them other than that they play certain roles specified in the story, and perhaps stand in certain relationships to one another. In effect what we know about them is implicitly defined in the story, couched in the vocabulary of O-terms; in effect the names ‘X’, ‘Y’, ‘Z’, and ‘W’ are T-terms.

From the story we can readily construct a Ramsey sentence in which the variables range over persons: there exist \( x, y, z, \) and \( w \), who . . . (and here a long conjunctive sentence spells out the story about the roles that the four persons played, all of which is couched in the O-vocabulary of the descriptive language used by investigators). It is also possible to construct four definite descriptions of the form: the unique person \( x \) such that \( x \) did so-and-so. Also names can be introduced for each person using one of the four definite descriptions for each. Finally, the world is such that the story in its Ramsey sentence form is satisfied by four people in the 4-tuple
\( \langle P_1, P_2, P_3, P_4 \rangle \). (Using Russell’s distinction, because of the story in which they play a role we know these persons by description; but we do not know them by direct acquaintance—or at least we do not know that we are acquainted with them.) Moreover, the story might be such that no other 4-tuple of persons satisfies the Ramsey sentence as well as the 4-tuple \( \langle P_1, P_2, P_3, P_4 \rangle \) does, or other 4-tuples satisfy it badly or not at all. In this way we can say that the story is uniquely satisfied (or has a unique realization). Finally, the four definite descriptions that can be constructed, and the names that can be introduced via the descriptions, will denote, in order, each of the persons of the 4-tuple \( \langle P_1, P_2, P_3, P_4 \rangle \).

The above example has analogies with the applications of the Ramsey–Lewis approach to scientific theories, such as the simple versions of theories of the electron or the classical atom already outlined. A growing body of reports of observations or experiments can be collected over time, all of which are expressed in the vocabulary of “outsider” O-terms. To this may also be conjoined some older well-established laws or even theory, again all expressed in “outsider” O-vocabulary. We need to allow that a given scientific theory can incorporate not only large amounts of observational information but also other laws and/or theories of science that are expressed using terms that are not to be defined within the given theory, and so are “outsider” terms. The given theory also contains “insider” T-terms, which are distinguished by the fact that they get their meaning in virtue of the role they play in the context of the given theory. Here the Ramsey–Lewis analysis of the theory will specify, along the lines already indicated, what it would be like for anything in the world to realize the given theory (if anything does). But it does this using only the old vocabulary of O-terms to tell a story about the roles that any realizer of the story will have to play.

Not all uses of the Canberra Plan in philosophy can be modeled in exactly this way. But the Ramsey–Lewis approach lays down a broad approach to methodology within philosophy that has wide application in posing problems and providing solutions. The essays in this volume investigate some of the different ways in which Canberra Planners might proceed and some of the problems they might face in realizing their program. In the detective story and in the illustrations from science, there is a growing body of information that can be readily identified. However, in some cases, there may be no such growing body of information to collect; rather, it might be at hand in our ordinary talk of matters relating to some domain of phenomena. A prime candidate here is our “folk” beliefs of some domain such as that of colors, or that of our folk psychology (in particular our beliefs and desires), or our folk morality about what is right or wrong. Such
beliefs may be common knowledge and known to be common; or they might not be commonly known and are implicit rather than explicit in much the same way as we have tacit knowledge of grammar.

In his essay in this volume Daniel Nolan sets out what he calls a Canberra Plan “two-step.” Suppose we wish to provide a philosophical analysis of the central concepts pertaining to some domain, X, such as causation, free will, color, morality, our psychology of beliefs and desires, or whatever. The first step will be to collect together the “platitudes” concerning the X to be analyzed. How platitudes about X are to be gathered is not always a straightforward matter. (It will be harder in the case of implicit knowledge, an issue raised in chapter 3.7) But in some cases they may simply be the large number of ordinary (or suitably refined?) beliefs we have about X, or the meanings implicit in our use of the central terms employed in talk about X, or the description of some paradigm cases of X, or all the truths about X (or what experts can agree are the truths about X), and the like. Exactly what counts as a “platitude” about X is one of the tasks that a Canberra Planner needs to address. There is also the issue of whether there is a sufficiently unified body of platitudes about X; and if not, since there is some degree of disunity, questions can arise as to how one might embark on the first step of the Canberra Plan.

If there is a sufficiently unified set of agreed platitudes about X, then there will emerge a theoretical role for the central notions describing the domain in which we are interested. Our agreed-upon platitudes about, say, color, yield what Lewis calls “the Moorean facts that constitute our folk psychophysics of color.” This results in a long conjunction of sentences upon which there is general agreement, such as ‘grass is green and normal perceivers in normal conditions have experiences of green caused in them & snow is white and normal perceivers in normal conditions have experiences of white caused in them & . . .’. This will contain terms of an “outsider” O-vocabulary; but the platitudes will also contain “insider” T-terms, the terms for the colors of objects and experiences, that play a role specified in the platitudes. This role can then be completely specified by using the Ramsey–Lewis analysis already outlined in which the T-terms are “Ramsey eliminated” by existentially quantified variables. But of course, what the variables range over is not thereby eliminated. So what in the world (or in our best theory of the world) do they range over? The second step in the Canberra Plan will be to discover what in the world, if anything, plays the roles so described; or, in a somewhat different vein, what our current best theories tell us there is in the world to serve as realizers of the theoretical roles specified in the platitudes. Here Canberra Planners will
commonly look to the deliverances of science for the realizers of these roles; they need to know what the world is like in order to “locate” the domain (for example, of colors) in whose analysis they were initially interested. (Whatever else may be involved, in the case of colors as surface reflectance properties there will at least be a certain quantum-chromodynamical account to uncover, and in the case of color experiences there will at least be an account based on patterns of neuronal firings in the visual cortex to uncover.)

The success of the search for realizers of the roles spelled out in the platitudes may well vary from case to case and in degree of satisfactoriness. It is not obvious that the platitudes, once the theoretical terms are “stripped out” along the lines of the Ramsey–Lewis analysis, will give a unique, fully satisfying fix on “something” in the world that plays the requisite roles (alternatively, a fix on items postulated in our current theories of the world). The very sciences to which an appeal is made might themselves be controversial. Or given our current sciences, perhaps nothing in them can serve as a perfect realizer of the specified role; but among the imperfect realizers there may be a unique best, imperfect realizer that can fill the specified role. Issues of ambiguity can arise if there are two or more best but imperfect realizers. And one might have to be eliminativist (as we have been in the case of phlogiston) if none of the imperfect realizers is satisfactory, or if there are no obvious realizers at all.

Most Canberra Planners in fact take it that in most cases if there are no physical realizers then there are no realizers simpliciter. This is because they are usually physicalists. One motivation for the style of analysis is to be able to locate in the physical world those things that are not obviously physical, whether they be minds, inflation, beauty, or rightness. But it is sometimes thought that the Canberra Plan is supposed to provide a methodological argument for a general physicalism. This is not so. In the case of the philosophy of mind there is a particular argument; it’s thought that the analysis reveals that mental states play causal roles, and it’s thought that we have a posteriori to believe that all causal roles are played by physical things. But this does not generalize. To the extent that Canberra Planners are physicalists in general, it is for the usual reasons of Ockhamism, induction from the success of physical explanation, and so forth. For physicalism is taken to be a contingent matter—even in the case of philosophy of mind it is entirely contingent that causal roles are played only by physical items, assuming that this is so.

Analyses of individual items reveal what has to be true of something for it to count as an instance of that which is being analyzed. In general there
is no way to be sure either that there is a physical item of which the analysis is true, or that only physical items meet the conditions. What the analysis does is tell you that, on the assumption that physicalism is true, what physical items, if any, satisfy the analysis. If there are none, then one can either become an eliminativist about the domain, or give up on physicalism. The overgeneralization probably comes from the claim that the analyzed items' existence is entailed by the way things are physically, or that there is an a priori supervenience of the analyzed domain on the physical. But it should be noticed that both of these only tell you that it is an a priori matter that a certain configuration of the physical is sufficient for the existence of the analyzed domain. This alone does not establish, of course, that the items in this domain are or must be physical; at best it establishes only that some are. Independent considerations must be used to argue for the former claim. So the analysis gives a kind of plausibility condition for physicalism. It tells us how much of our commonsense ontology would be vindicated if physicalism were true.

Concerning realization, there is no clear methodological way of proceeding to determine what the realizers are, whether perfect or best imperfect. Concerning best imperfect realizers, Lewis does tell us that “the notion of a near realization is hard to analyze but easy to understand” (Lewis 1972 [in 1999], 253). A number of other constraints can come to apply to determine what in the world can serve as a realizer other than fit with the role specified by platitudes. Here is one example. Suppose A and B are equally good but imperfect realizers of a theory but that A is a more “natural” property while B is less natural and more “gerrymandered”; then the recommendation is to choose A over B.¹⁰ Here an extrascientific consideration from metaphysics come to play a role in determining realizers that may not be best fit. Another way to resolve problems of best fit may be to return to the original set of platitudes and consider ways in which each might be differentially treated as more or less significant, thereby giving a weighted set of platitudes to be subject to Ramsey–Lewis regimentation.

Suppose that the second step is completed and satisfactory realizers of our platitudes have been obtained (for example, our platitudes about color come into a good relationship with the claims of science, so that color has a location within science). Then, one hopes, what realizes the roles will end up casting a more perspicuous light on the domain under investigation (color) than was available at the beginning before we embarked on the first step. The Canberra Plan “two-step” sets an agenda of analysis combined with a penchant for the naturalistic, given the involvement of science.
Such is the broad sweep of the Canberra Plan; but there are devils in the details.

Here is one important detail to conclude the introduction, the possibility of circularity. Canberra Planners are typically descriptivists. A Ramsey sentence is a description that picks out in the world whatever it describes, if anything. This description may say that what it takes be a certain kind of entity is to interact causally in a certain way, or structurally in a certain way. Description is of course a semantic relation of a certain kind, and in order to be in the business at all we need to be able to use descriptions. A question arises, then, about whether it is possible to give an analysis of representation. Descriptions are representations, and so to perform a Canberra Plan analysis we need to use our power to represent. But if what we are representing is the nature of representation, there is a faint whiff of circularity. Certainly it would be no good if the Ramsey sentence itself mentioned representation. To know that representation is that thing in the world which is related in various ways by the representation relation is not to know much. In addition, it would violate the principle that the analysis be solely in terms of the old vocabulary.

We do not, however, need to use an analysis of representation in order to give an account of representation. We need merely to represent. Using the power of representation does not involve having a theory of it, necessarily. So if there are platitudes about representation that tell us that representation has certain structural features, and it turns out that there are such things in the world—covariation relations, functional connections, analog isomorphisms, information relations, or whatever your favorite reductive account is—then we have a solution to the location problem for representation. Just so long as we are beings who can represent in language and thought (which though controversial in some quarters is taken for granted here), and just so long as the platitudes do not mention representation, then all is well. Of course once we have our solution to the location problem we had better see that it vindicates itself. In other words, when we have an account of representation, we need to see if the relation in the world that we have picked out is one that obtains between itself and the (perhaps beliefs about the) Ramsey sentence we began with. Suppose that we find the relation in the world that is picked out by the descriptive Ramsey sentence is, for example, causal informational covariation. We would then need to see that this relation holds between beliefs whose content is given by the Ramsey sentence and the general relation of causal informational covariation. But this is simply an adequacy condition on the theory. It is not how we establish the theory in the first place. So, just as
long as we are representers and have the skill of representing, and do not mention representation in the Ramsey sentence, we doubt that there is any danger of circularity here at all.

2 The Essays in This Collection

The papers in this collection are divided into three parts addressing issues in mind, metaphysics, and normativity that arise in implementing the Canberra Plan in these domains of philosophy. The seven papers in Part I deal with some aspect of the application of the program to issues to do with the philosophy of mind, semantics, concepts, and the Plan’s a priori character.

The first chapter in Part I by David Braddon-Mitchell, “Naturalistic Analysis and the A Priori,” sets the scene for some of these problems. He begins by considering the ways in which the distinction between O-terms and T-terms was initially drawn by positivists. Not only is their way of drawing the distinction a failure, but so also are the kinds of analysis they attempted based on the distinction. The Canberra Plan importantly distances itself from these positivist failings; it also distances itself from the way in which they attempted to incorporate naturalistic considerations into their analyses. He then turns to two matters concerning the a priori nature of the Canberra Plan. The first is an objection and asks: how is metaphysics to be conceived, if it is taken to be neither empirical nor a priori? If it reveals a priori necessary truths, then competing false hypotheses will be intensionally equivalent, and thus analytically equivalent given some usual presuppositions of the Canberra Plan, such as its reliance on two-dimensional semantics. So there is a need for a theory of the hyperintentional. The second matter deals with a further question about the a priori, a question also raised in the next chapter by Stich and Mason in their correspondence with Jackson. The question is: if it is matters of our dispositions to behave about which we are extremely fallible that determine our meanings, how can analysis be a priori? Finding this out is a difficult a posteriori matter. Braddon-Mitchell’s response is that in these cases what matters is our judgments about what to do and say should our dispositions turn out to be other than we predict from the armchair.

One of the key ideas shared by Canberra Planners is that of tacit knowledge. The analytical functionalist about the mind, to take the classic example, thinks that there is a theory of the mind—folk psychology—that we believe and which would, in analysis, be systematized. But the functionalist does not think that we believe it explicitly; we cannot write down
such a theory, and we do not have access to sentences that systematize it. The dialogue between Jackson on the one hand, and Stich and Mason on the other, addresses this problem. Stich and Mason have a challenge. If the theory is tacit, it is hard to see how this could be unless it was encoded in the brain at a level to which we do not have good introspective access. In that case it might give the right explanation of certain bits of behavior, but it would in no good sense count as being believed—and certainly not count as something to which there is any a priori access. If, however, the theory is encoded in the brain, it must be a learned theory about the behavioral patterns of others: a theory that is a generalization about behavior. But this looks like a theory that one might generate if one were doing a certain kind of empirical work, but once again not one that is believed by most people, and certainly not believed a priori. One might construct such a theory, but one would not be doing analysis. In the dialogue the parties try to sort out what is meant by tacit, what is meant by pattern, and how a tacit theory could count as a belief. Jackson’s answer is roughly that the theory is expressed in an agent’s pattern of judgments and behaviors. Insofar as we have access to these we have access to the theory. The theory is precisely not one whose level of detail goes beyond what is expressed in that pattern—that would indeed be an implementational theory of how cognition is performed, which is a posteriori and not a subject of belief in any but cognitive sciences. Given the functionalist account of the content of belief that Jackson subscribes to, the dispositions that are expressed in these patterns determine the content of the beliefs.

Suppose that the existing vocabulary refers to a set of entities and properties that exhaustively but only contingently characterize the actual world, and suppose our analyses give accounts of the roles that, as it happens, are played by those contingently existing properties and particulars. The Canberra Plan approach to philosophy of mind has this feature. Our analyses of the mental are in terms of roles that are played by physical states. But physicalism, if it is true, is contingently true. So those roles could be played by nonphysical things; if the right functional roles are played by, say, ectoplasm in some world, then (according to functionalism) that ectoplasmic being has a mind. What this means, of course, is that which unites all possible mental beings is a property that can be possessed by both physical and nonphysical things, even if physicalism is in fact true. Frank Jackson’s chapter, “A Priori Biconditionals and Metaphysics,” explores this thought and what it means for Canberra Plan analyses. The idea he defends is that what makes a world physical is that it contains only physical things, the aggregations of which a priori entail that a certain
pattern is instantiated. The idea is that in coming up with an analysis we do two things. We first give the content of a concept in a way that is neutral, so we can see how many different worlds containing different sorts of underlying ontology could make it true that the patterns of which we have an a priori grasp are instantiated. Second, we are in a position to tell, should we know a posteriori what our world contains, whether it is a world that instantiates one of these patterns. This last capacity is itself a priori—for it is an a priori matter what patterns are entailed by what distributions of different basic properties and particulars, even though it is an a posteriori matter which distribution our world contains.

One important question raised by the Canberra Plan methodology is what to say in the event that it turns out that there is nothing of which the a priori theory is true. Of course one option is to simply become eliminativists about the area in question. But sometimes that seems unacceptable, and indeed there are things in the world of which a similar theory is true. This can be used to explain away some difficult problems. A case in point is some of the arguments against physicalism in the philosophy of mind. These purport to show that there are things in an a priori theory of consciousness or qualia that could not be true of any physical states. A diagnosis of these arguments that has been very influential is that it is indeed the case that the world contains nothing that satisfies these a priori theories about consciousness, in particular the idea that the (so called by David Lewis) Identification Hypothesis is true of qualia. But nonetheless the idea that we are conscious, and have qualitative experience, is explained by the idea that we have a replacement conception of consciousness or qualia. There are physical states in the world of which this replacement conception is true. So we give up the original theory, and revise our beliefs to the alternative theory, and call the things in the world of which it is true ‘experiences’. Daniel Stoljar’s chapter, “The Argument from Revelation,” challenges this move. After clarifying just what the Identification Hypothesis is, he argues that it is not in fact presupposed by common sense, or even in fact by some of the antiphysicalist arguments. If Stoljar is right, then the Canberra Plan does not provide a straightforward way out of these arguments, and nor does the case of qualia play a role in vindicating a methodological template for solving a certain kind of philosophical problem.

Fred Kroon, in his chapter “Names, Plans, and Descriptions,” takes the Canberra Plan to be a family of distinct doctrines, united by a confidence in a broadly physicalist worldview and the ability of a priori philosophizing to support and elucidate the way our ordinary talk and thought really is
talk and thought about the world as physics reveals it to be. The essay addresses the nature of a priori philosophizing distinctive of the Canberra Plan approach, and considers some of the challenges that this faces. In particular, it considers a recent challenge by Scott Soames to the sort of a priori descriptivism defended in the main by Frank Jackson. Kroon argues that Soames’s challenge is best answered by incorporating the idea that our referential practices are appropriately motivated, and that an appeal to such motivation explanatorily underwrites the relevance of the sort of properties uncovered by a priori descriptivism.

What is the connection between the analysis of words and the analysis of concepts? Kingsbury and McKeown-Green’s chapter, “Jackson’s Armchair: The Only Chair in Town?,” addresses this question, after replying to a number of objections to Jackson’s version of Canberra Plan reductive analysis. Their contention is that Jackson conceives of conceptual analysis almost exclusively as the analysis of lexical items, and that this impoverishes the program. The thought is that how we use words is only a small part of our cognitive and conceptual repertoire. In the central case of belief, for example, the situations in which we will attribute beliefs to others by agreeing that the word ‘belief’ be used of them constitute only a small part of a practice of explaining, predicting, and interacting which we might explain by our possession of the concept of belief. The right kind of analysis, they contend, is done by investigating this richer pattern of thought and behavior. Language may come into play afterward, when we give names for the concepts we find to be structuring our thought. The version of analysis they defend not only has the benefit of transcending the contingencies of the words we use, it also insulates the debate about our a priori access to our conceptual competence from the debate about whether our semantic competence with words is modularized, and thus not accessible to introspection.

In their chapter “Is Semantics in the Plan?” Peter Menzies and Huw Price raise a question about the two-step Canberra Plan. The first step is that in which a number of “platitudes” from, say, folk psychology are collected about mental terms such as ‘belief’, ‘desire’ and the like. From these is generated the appropriate Ramsey sentence, ‘whatever it is that plays the R causal role’, or the appropriate Lewis description, ‘the unique so-and-so that plays the R causal role’. They point out that such a first step does not employ in an essential way any nondeflationary semantic notions. As a preliminary to the second step, one can ask: what is it in the world that plays the R causal role? One trivial answer is: “of course it is the mental states of belief and desire.” While not rejecting this, the Canberra Planner
would not regard it as a complete answer. What the first step picks out (originally mental entities) is ‘whatever is the occupant of the R causal role’. The second step of the Canberra Plan turns on the closure of the physical and tells us that the occupant can be fully studied in terms of the physical sciences. Here causation plays a vitally important role in making the a posteriori identification of the mental with the physical. However, the Canberra Plan has been generalized to cases in which there are no causal roles to which to appeal; rather causal roles are replaced by the more omnibus functional role that a T-term can play (as in the application of the Canberra Plan to moral functionalism in which functional roles are not normally causal roles). In the first step of the application of the Canberra Plan platitudes of folk morality are collected, and then it is asked “what (unique) thing plays the F functional role?” One trivial answer is, of course, moral properties such as being right, or being good, and so on. But this again is not the desired final answer in terms of descriptive properties only. So how is the second step of the Canberra Plan to be carried out? Menzies and Price Menzies argue that it cannot be along the lines of the first case in which nondeflationary semantics plays no essential role; rather they see an appeal to nondeflationary semantics such as reference as an essential ingredient in uncovering whatever it is that plays the functional role (other than the trivial appeal to moral properties). Given this they argue that there are two descendants of the original version of the Canberra Plan that can be found in the work of David Lewis. The first is one that appeals to causal roles without an appeal to nondeflationary semantics; the second is one that appeals to functional roles (that are not typically causal), in which there is a heavy-duty employment of semantic notions.

The four essays of Part II address matters of metaphysics in relation to the Canberra Plan. The first two deal with the claim that Ramification leads to the multiple realizability of any putative final theory of the world, and so to an irredeemable ignorance about the world’s fundamental properties. The third raises a problem of “derivational deficiency” for physicalism, and the fourth shows how the platitudes required by the Canberra Plan can be a resource for metaphysics.

David Lewis, in his seminal 1970 “How to Define Theoretical Terms” provided a novel way for fixing the denotation of T-terms that lies at the heart of the Canberra Plan. Important to his view was the idea that theories can have a unique realization (by an \( n \)-tuple of entities) and that T-terms denote each of the unique realizers (in the order in which they appear in the \( n \)-tuple). This idea can be modified to allow that if there were no unique perfect realizers then there could be imperfect best realizers. It can also be
modified to allow for indeterminacy of denotation where two or more \( n \)-tuples realize the theory equally well; such indeterminacy can be resolved with further scientific advances (Lewis 1994 [in 1999], 301). In one of his last papers, “Ramseyan Humility” (published here for the first time), Lewis departs from the view that there is a unique realization for our theories. First, he shows that if there is a final theory, then its Ramsified version will be realized by an \( n \)-tuple of “fundamental” properties, which provide a full inventory of the properties at work in nature. However, he goes on to argue that even though there is a unique actual realization, there are also many possible realizations of the theory. Can we ever tell which of the possible realizations is the actual realization? No. The several strands of his argument for this position call on a number of important theses in metaphysics. The “permutation” strand draws on the notions of combinatorialism and quidditism; other strands of the argument embrace the possibility that there are “idler” and alien properties to take into account. Whichever strand of the argument we follow, they all lead to the same conclusion: given the different possible realizations of the Ramsified theory, we have no way of telling which realization is the actual realization. Ramseyan Humility is our irredeemable ignorance about the identities of the actual fundamental properties that realize the final theory. In setting out his complex argument Lewis is fully aware of the ways in which it might be countered; but he sets out a consistent set of metaphysical and epistemic doctrines that lead from Ramsification to Humility.

Lewis’s argument has already excited commentary that queries the skeptical conclusion of Humility. The chapter that follows by Dustin Locke, “A Partial Defense of Ramseyan Humility,” sets out to defend Lewis’s arguments from some of its critics who would reject one or another of Lewis’s premises. First, he gives a representation of the metaphysical part of Lewis’s argument, adding some extra theses to make it more explicit. He then develops the epistemic part of Lewis’s argument, discussing the critics who all (implicitly or explicitly) query Lewis’s implicit assumption of his own account of knowledge. What Locke goes on to show is that the conclusion of Humility follows not only on Lewis’s own epistemic and semantic assumptions but also on the alternative assumptions of the critics. Locke brings out aspects of Lewis’s argument that turn on two-dimensional semantics to resolve issues about exactly what proposition it is that Humility expresses. Once this is clear Locke then shows that Humility is not to be identified with standard versions of skepticism through the investigation of a number of well-known responses to classical skepticism; standard appeals to abduction, or to some version of anticlosure, and the like, fail
to deal adequately with Lewis’s argument. However, what Locke and the critics he investigates do agree on is that Humility is not an ominous doctrine of massive ignorance but a benign form of ignorance. Classical skepticism tells us that some of the things we thought we knew we do not know; the ignorance that Humility establishes is ignorance of something we never thought we would know anyway, given our lack of the supposed final theory.

Physicalism is an important accompaniment to the Canberra Plan. In his “Physicalism without Pop-out” Philip Pettit endorses a version of physicalism that says that there is an a priori entailment from the way things are physically to the way they are in other respects, such as the psychological or social. But there is a general problem with this; though we may believe that there is such an a priori derivation of the psychological from the physical, we suffer from “derivational deficiency” in that we are hardly in a position to make the derivation or even to see how the derivation might go. To illustrate the deficiency he begins with an analogy of a system of dots with specific coordinates on a grid and the shapes that they may give rise to, such as the following: a straight line, an S-shape, or a picture of a person. In each case, can we expect an inferential “pop-out” concerning each of the shapes, given our a priori knowledge of design specifications (viz., that any system of coordinates of dots in pattern P generates a particular shape), and the empirical information that such-and-such a particular array of dots is in that pattern? Here there is a ready and correct conclusion to be drawn, namely, that such-and-such pattern of dots generates the particular shape. But although there is commonly derivational pop-out in the case of the straight line that is also phenomenologically satisfying, this is less so in the case of the S-shape and may not be so at all in the case of the picture of a person. Given these cases Pettit distinguishes two kinds of derivational deficiency, shallow and deep. On the basis of the analogy he shows how both kinds of derivational deficiency carry over to physicalism’s claim to derive a priori the psychological in the case of representational states. We may well know a priori that the “pattern of dots” of neuronal states generates a “shape,” that is, representational states, and have empirical information to the effect that a particular person’s neuronal states are of that “pattern.” But given such information there is no ready “informational pop-out,” in much the same way as in the case of the dots there is no ready “visual pop-out.” Pettit goes on to show that shallow derivational deficiency arises for nonrecursive representational states (i.e., there is a representation of the environment for a subject without representation that the environment is so represented); in
contrast, there is deep derivational deficiency for recursive representational states (i.e., states in which not only is there is representation that the environment is so represented, but the subject can also adjust and act accordingly). His conclusion is that the physicalist is deprived, for the time being at least but not irredeemably so, of the requisite kind of inferential pop-out.

In his chapter “Platitudes and Metaphysics,” Daniel Nolan considers the ways in which “platitudes analysis” has become an increasingly popular technique in philosophy exemplified by the Canberra Plan. A set of widely accepted claims about a given subject matter are collected, adjustments are made to the body of claims, and these are taken to specify a “role” for the phenomenon in question. One of the best-known examples is analytic functionalism about mental states, where platitudes about belief, desire, intention, and the like are together taken to give us a “role” for states to fill if they are to count as mental states. The next task is to look to our best theory of the world to see whether this role is satisfied, if at all. Unfortunately, platitudes analysis, so characterized, does not seem to help when we are doing fundamental metaphysics, that is, when we want to know what, at base, our world is like (and not merely where things such as, for example, the mental would be found in an already specified ontology). Despite this, platitudes analysis, properly understood, does have the materials to help us answer questions in fundamental metaphysics as well. Nolan explores three different ways in which this is so.

The two essays included in Part III raise matters concerning normativity in relation to the Canberra Plan. The first considers how the normativity of rationality can be naturalized; the second considers problems for the Canberra Plan raised by the normativity of ethics and the possibility of moral disagreement.

In “Naturalizing Normativity,” Mark Colyvan discusses the problem of providing an account of the normative force of theories of rationality. The theories considered are theories of rational inference, rational belief, and rational decision, that is, theories of logic, probability theory, and decision theory. He provides a naturalistic account of the normativity of these theories that is not viciously circular. But he also points out that the account does have its limitations: it delivers a defeasible account of rationality. On this view theories of rational inference, belief, and decision are not a priori; rather they are a posteriori and may change over time. Finally he compares his approach with another which emerges from the Ramsey–Lewis approach to defining theoretical terms that lies at the core of the Canberra Plan.
One of the paradigm applications of the Canberra Plan is to the domain of ethics, in particular Jackson’s and Pettit’s moral functionalism in which our ordinary folk opinion on moral matters plays an important role in determining which morally evaluative properties are which descriptive properties. But is our ordinary folk opinion on moral matters sufficiently unified to play the role required of it, or is it sufficiently disunified so that it cannot play such a role? Answering this question is the main task of Denis Robinson’s chapter, “Moral Functionalism, Ethical Quasi-Relativism, and the Canberra Plan.” His account of ethical quasi-relativism is not the simple relativism due to differing agents and/or their standpoints, since these do not give rise to genuine disagreements; rather it is a version of relativism that admits the legitimacy of different concepts of right, wrong, permissible, and the like, yet nevertheless acknowledges disagreements about moral matters as bona fide even when they turn on those conceptual differences. The difference between genuine disagreements and those that are not genuine is illustrated by a sequence of small dialogues. The upshot of Robinson’s discussion is that the fact of pervasive moral disagreement and controversy raises difficulties for moral functionalism. To meet these difficulties it has been suggested that the idea of ordinary folk morality can be replaced by a version of mature folk morality that results from subjecting our ordinary folk morality to critical debate and scrutiny in which there is convergence of moral opinion in the long run if not the short. But this maneuver is unavailing and does not remove the possibility of bedrock disagreement. As a result ethical quasi-relativism takes seriously the idea that there is no unique property to be determined, in Canberra Plan fashion, by either ordinary or mature folk morality. The paper concludes with suggestions about how to model ethical disagreements; but they would be small comfort for Canberra Planners.

Notes

1. For some who have used the expression “Canberra Plan” in their published work, see: section 1.1 and footnotes 3 and 4 of the extend reprint of Lewis’s “Causation as Influence” in Collins, Hall, and Paul 2004, 75–106; Nolan 2005, 223, 237.

2. See the end of Carnap 1966, chapter 26, for his agnosticism about the realist-instrumentalist controversy; he deploys his understanding of the Ramsey sentence in an attempt to show that the source of the controversy is mainly linguistic.

3. See Psillos 2000 for an interesting story of how Carnap developed his account of Ramsey’s theory under the influence of Hempel and his rediscovery of Ramsey’s work that he had much early read but had forgotten. For Carnap’s Ramseyan
approach, see his reply to Hempel in Carnap 1963, section 24, 958–996 and Carnap

4. Here we pass over the various ways in which Lewis’s approach is similar to, or
different from, Ramsey’s. One difference is the way in which the T-terms that are
not names but predicates or functors can be put in name position so that they can
all be replaced by variables of the same style. This need not necessarily be so in the
case of all deployments of the Ramsey sentence; see Lewis 1970, section I.

5. If there is just one denoting definite description, then it will denote the single
entity A; but if there are two denoting definite descriptions then each will denote,
in order, the members of the pair ⟨A, B⟩. In the general case of n descriptions there
will be an n-tuple of entities ⟨A, B, . . . , N⟩ such that the first description denotes A,
the second B, and so on. On the basis of each description a term can be introduced
which denotes the entity that the description picks out; see Lewis 1970, section
IV.

6. Later Lewis modified his position allowing that where more than one entity
answers to a description then the description does not fail to denote but denotes
ambiguously. See his 1984 paper “Putnam’s Paradox” as reprinted in Lewis 1999,
59 where he talks of indeterminacy of reference; see also Lewis 1994 (in 1999), 301,
where he talks of ambiguity of reference.

7. See also Lewis 1994 (in 1999), 298, on tacit knowledge and his claim in the
footnote that “eliciting the general principles of folk psychology is no mere matter
of gathering platitudes.”

8. See section II of “Naming the Colors” in Lewis 1999, chapter 20, 332–358; this
discusses the Moorean facts of the psychophysics of color. In this chapter Lewis also
discusses objections to his Ramsey–Lewis approach to issues about color properties
and color experiences; in this respect also see the essay by Stoljar in this volume.

9. This way of expressing the matter and an excellent account of Lewis’s method

For the role that natural properties can play as realizers of theory see his paper
“Putnam’s Paradox” in Lewis 1999, chapter 2, especially the appeal to elite properties
in the section entitled “What Might the Saving Constraint Be?,” 64–68.
I Mind, Concepts, and Theories
2 Naturalistic Analysis and the A Priori

David Braddon-Mitchell

1 Naturalism and Analysis

The relationship between naturalism and analysis has been a stormy and varied one. It’s gone from almost conjugal closeness to bitter animosity and back to tentative signs of reconciliation. The relationship begins in recent times in the positivism of the early twentieth century. Here it seemed the job of philosophy as performing conceptual analysis was integral to the naturalistic and scientific worldview. The positivist partition of all sentences into observation sentences and theoretical sentences left room only for two kinds of kosher intellectual endeavor: science, which collects and systematizes observation sentences, and analysis, which systematizes the analytic rules that govern the use of all the other sentences.

So the philosopher as analyst was the handmaiden of science. The philosopher noted the uses of theoretical terms, and made them explicit and consistent. Thus philosophically cleansed, they could better carry out their job of organizing observations. Analysis typically did not consist in gathering together explicitly held opinions about theoretical sentences, but rather in analyzing what must be tacitly held, based on evidence from the use of terms. Thus behaviorism as a theory of mind emerged from positivism, not through introspecting on the meaning of mental state terms, but through examining how and under what conditions mental predicates are applied.

But the positivist project failed. No one could come up with a clear and purely logical criterion for distinguishing between observation sentences and theoretical sentences. This was part of a widespread failure of the grand scheme to use purely logical resources to define much of what is foundational to the scientific and philosophical enterprise.
And, of course, then came Quine. Quine, in attacking the analytic–synthetic distinction, undermined the basis for the harmonious relationship between analysis and naturalism. By insisting that putatively analytic sentences are simply sentences subject to the tribunal of experience—albeit with more solid alibis than many others—he created a climate in which analysis seemed out of kilter with the naturalistic worldview.

The job that Quine began, Kripke completed. Kripke taught a generation of philosophers that analysis was not up to either of the two tasks it had often been assigned. It certainly didn’t reveal the essences of things, and it didn’t even reveal the meanings of terms. All analysis did was make explicit fallible beliefs about things. Famously, we all began to think that everything we believed about something—say, water—could turn out to be false. And despite this, we could still have thoughts about water, and our words could still refer to water. The trick was to see that not only were all the essential properties of water ones that could be discovered a posteriori, but the nature of meaning and reference itself was given by the best scientific theory. The post-Kripkean orthodoxy became that matters of reference were doubly a posteriori. On the one hand we have to develop an a posteriori theory of meaning, and on the other hand that theory will then tell us which a posteriori features of the causal histories of words and concepts are relevant for settling matters of reference.

The decline of the relationship between analysis and naturalism was thus close to complete. Explicitly analytic methods became the explicit methodology only of those philosophers who eschewed much connection with the natural sciences, or even who were antagonistic to the claims of naturalists that the scientific endeavor might exhaustively catalog the basic furniture of the universe. Analysis began to acquire one of two reputations. On the one hand, there was sterile cataloging of pointless folk wisdom—such as articles analyzing the concept VEHICLE, wondering whether something could be a vehicle without wheels. This seemed like trivial lexicography. On the other hand, there was metaphysically loaded analysis, in which ontological conclusions were established by holding fixed pieces of folk wisdom—such as attempts to refute general relativity by holding fixed allegedly conceptual truths, such the idea that motion is intrinsic to moving things, or that there is an objective present. This seemed to many a case of philosophy seriously overreaching itself.

At the same time, naturalistically minded philosophers were looking for accounts of things obtained in ways free of the taint of analysis. Very often the methodology took its cue from the so-called new theory of reference; the search was on for natural kinds causally and historically related to the
uses of the terms associated with whatever was in need of philosophical explication—genes, causation (the conserved quantity theory springs to mind), persistence, or even democracy.

Something was odd in all of this. The naturalists may have railed against analysis, or armchair wisdom, but thought experiments continued to play a key role, albeit sometimes suppressed, in much of what they did. Indeed the ball was got rolling, as it were, by the Twin Earth thought experiment. Thought experiments like this, rather than any a posteriori investigation, were what revealed that intrinsic natures mattered for the essence of natural kinds—while of course not revealing what those intrinsic natures were.

2 The Canberra Plan

The so-called Canberra Plan—the style of philosophy influenced by David Lewis and Frank Jackson, whose practitioners often seemed to pass through the Philosophy Programme at the Research School of Social Sciences at ANU during the 1990s—promised a reconciliation between naturalism and analysis. The key idea was to abandon the requirement to analyze the distinction between observation sentences and theoretical sentences using logic. Instead, taking its lead from Lewis’s “How to Define Theoretical Terms” (1970), a more easily satisfiable goal was adopted. The job of philosophy is to analyze, but analysis consists in making Ramsey sentences out of existing vocabulary so as to define new terms. The new terms that are thus defined are the objects of analysis; a priori investigation tells us what is true of them using existing vocabulary (or more clearly understood concepts, in the nonlinguistic case). Then we get to look and see what things in the world play the roles defined by these sentences, and if there is some unique particular or kind that does, we have discovered the nature of what we set out to analyze.

What made this seem more attractive than positivism was its less ambitious scope. Exactly because the idea is to analyze contested language (or concepts) in terms of uncontested or existing ones, many of the problems are brushed under the carpet. Major divisions of philosophical opinion will be reducible to questions about what is or is not contested.

2.1 How Deflationary Is the Canberra Plan?

An example: suppose that you were attracted to a strong kind of positivism according to which all the basic uncontested claims are claims about sense data. Then a version of the Canberra Plan semantics would simply be that
there is a lot of existing, basic vocabulary that has as its content claims about sense data. Theoretical language would then Ramsify over these claims, providing Lewis–style analyses of everything else in terms of the interrelation of claims about sense data.

But this, of course, is unattractive to most, and the point is to avoid classical positivism. Sometimes philosophers in this tradition are tempted to a kind of naturalistic weak positivism, where the uncontested (for philosophical purposes) vocabulary is all and only the findings of the mature natural sciences. Then we get to give analyses of everything else in terms of these, by finding whether there is some entity describable in this language that plays the roles described in the analysis. Such a view would be unfriendly to first-order metaphysics as it is usually conceived. The basic vocabulary would have all and only terms found in mature natural sciences; and the theoretical terms would specify only roles that must be played by things whose nature is given by the mature natural sciences. Thus, for example, in the classic case of philosophy of mind we might be able to specify the belief role in a Ramsey sentence, and then the only allowed move would be to see if there is something in our scientific ontology that plays that role (or near enough), in which case we have vindicated beliefs, or else we discover there is no such thing, in which case we have discovered that eliminativism about belief is true.

But this leaves little role for traditional metaphysical questions. A dualist in the philosophy of mind, for example, might think that the whole procedure begs the question. For dualists think that beliefs might be some kind of nonphysical state or property—they think that there are additional basic properties in the universe.

How can we make sense of this disagreement? There are, I think, two ways to potentially loosen up the almost positivist vision of the Canberra Plan. The first, and closest to positivism, is to take it that whether there are extra properties beyond the basic physical ones is itself a question of first-order science. Something like the unity of science arguments for monism, and the inductive argument (all previous problems have been solved on a physicalistic basis), might be instances of arguments that could be thought to be internal to scientific method. Certainly there is a strand of physicalistic thought that takes physicalism to be a hypothesis within naturalism. On this reading, to be a physicalist is to accept that there are complete sufficient causal explanations for all events that are part of an integrated physical science. On the other hand, to discover empirically that dualism was true in the philosophy of mind would be to find evidence of special causal mechanisms that underlie mental behavior which do not
reduce to or supervene on the laws of nature that govern the rest of the world. Thus the discovery of special extra mental properties would be a discovery within science. It would be the discovery that it was at a fundamental level less unified than we might have hoped, and that the border between the disunified elements is the border between the mind–brain and the rest of nature.\(^1\)

The job of the philosopher would then be twofold; the first would be to specify the roles of mental state terms or whatever concepts she is working with—doing the analysis—and then second to see whether these roles are played only by entities in the widely integrated (and hence physical) part of the scientific ontology, the narrow exceptional part of the ontology, both, or neither.

The flavor of this method of accommodating metaphysical questions is consonant with an all-embracing naturalism that characterizes a number of strands of contemporary philosophy. But it casts a useful light on it. Post-Quinean naturalism, perhaps under the influence of Kripke, often tries to do two jobs at once—settle the ontology, and taxonomize the ontology. Thinking of things in a way that Daniel Nolan (this volume), for example, has called the Canberra plan two-step helps to restore some order. I’ll explain what I mean using the previous example. I suggested that on this strongly naturalist reading of how to do metaphysics on the Canberra Plan, the science first settled whether the ontology was dualistic or not, and then we looked at our analyses to see whether some role—such as the belief role, perception role, the causation role, and so on—was played by physical or nonphysical properties. But that can’t be quite right, because it assumes a prior analysis of “physical”—one according to which if there is a lack of integration of one part of fundamental ontology coupled with the borders being around the mind–brain, then that is enough to qualify that smaller part of the basic furniture of the universe as nonphysical. Better, however, that the basic ontology be crafted in neutral terms—such as simply that there is (or is not) a breakdown of the unity of nature around the borders of the mind–brain. Then we can do an analysis of “physical” to see if that distinction expressed in neutral terms answers to enough of our understanding of the physical–nonphysical distinction. Similarly, those who think that whether three-dimensionalism or four-dimensionalism is the right account of space, time, and persistence is an empirical matter might nevertheless think that what we want is an account of the way the world is in physical terms that is couched in neutral terms. Only then should we analyze concepts like PERSIST to see if what this neutrally described ontology provides is a good enough deserver to vindicate persistence.
So that’s one way to put metaphysics back in the Canberra Plan: it is to adopt it as a framework for explaining what people are doing when they engage in the strong naturalistic program of finding out a posteriori metaphysical facts. Perhaps it’s Quine with added analyticities. Perhaps it’s exactly what J. J. C. Smart was doing in his metaphysics and the philosophy of science—the “topic neutral analysis” being the analytic component, and the naturalism in looking for the things of which the analysis is true.

2.2 A Priori Metaphysics

However, this is not a way of looking at things that will appeal to some metaphysicians. Fortunately it’s also not the only way we can get the a priori back into the Canberra Plan. The second way of accommodating metaphysics is that it proceeds in large part a priori, in virtue of exposing hidden contradictions or inconsistencies. Some physicalists think that this is what they are doing in arguing against dualism, and some dualists likewise. Some philosophers of time think, for example, that the A series is contradictory, via a descendant of McTaggart’s argument.

But this conception of metaphysics is also one that the Canberra Planner can embrace. In gathering the platitudes and doing something like an analysis of what it takes to be a mental state, for example, it might turn out that there are features that seem as if they are important but which are nonetheless incoherent. So suppose in gathering together what’s important about our concept of qualitative experience the doctrine that David Lewis called the identification hypothesis turns out to be fairly central. That is, roughly, the idea that experience is unique in that when we have experiences, we learn their intrinsic essential properties. Physicalists about the mind might doubt this for a number of reasons; for example, that it is in conflict with physicalism, which is independently plausible (roughly Lewis’s own view in Lewis 1995). But one option is to think that it is incoherent in some way, that somehow experiencing something and it revealing its intrinsic essential nature are in logical tension (perhaps because of a metarepresentational view of awareness). I do not defend such a move here; I just use it as an example of this second way in which we might get metaphysics back into the Canberra Plan: by ruling out various philosophical hypotheses a priori. On this conception, the job of getting the platitudes together and Ramsifying will reveal some contradictory Ramsey sentences, and at the end of the process of Ramsifying we will have made metaphysical discoveries of a sort, insofar as we will have eliminated many inconsistent roles that cannot be played by anything. Only then
would we be ready to move on to find out if the coherent roles are in fact played in the actual world.

2.3 Two Challenges for the A Priori
So far, so good. We have a model according to which the Canberra Plan uses a priori methods to discover the content of our concepts, and to eliminate those concepts that are incoherent. But there are at least two major challenges for this conception of the role of a priori analysis. The first is that the Canberra Plan makes liberal use of the two-dimensional framework for explaining the content of words or concepts. This framework is not fine grained: it attributes the same content to all the a priori necessities or impossibilities. So the Canberra Plan will need an account of hyperintensional content, to explain what the difference is between the various hypotheses that we try to rule out in the a priori part of our enquiry. The second challenge is that the Canberra Plan attributes content based in part on behavioral and dispositional facts about agents that might not be a priori accessible. If so, the status of the analysis as a priori (rather than a kind of a posteriori cognitive science) might be in doubt.

These two challenges in different ways strike at the heart of the rehabilitation of a priori analysis. The remainder of this essay addresses them.

3 The Canberra Plan and Hyperintensionality
On the conception of metaphysics we have been considering, it must be that when we consider various contrary metaphysical hypotheses, we know that at most one of them in fact has ordinary possible worlds content, for at most one of them is coherent. So such a conception of metaphysics on the Canberra Plan urgently needs an account of hyperintensionality to explain what we grasp when in our ignorance we contemplate and appear to grasp the content of what seem to be multiple contrary coherent hypotheses.

But here we come to the crunch. What is it that we grasp when we appear to grasp what seem to be different metaphysical hypotheses, and we appear to understand the difference between them? If they are metaphysical hypotheses that go beyond the empirical, we should think that one of these three conditions holds:

(1) All of the hypotheses are internally contradictory, and if we appear to grasp differences between them, then these are not intensional differences—they are not different hypotheses about how things might be.
(2) One of the hypotheses is coherent, the rest incoherent. In this case there is an intensional difference between the true view and the rest (but not between the true view and all the other true metaphysical views).

(3) More than one, perhaps all, of the hypotheses are coherent. The coherent ones, if they are strongly empirically equivalent,2 will be intensionally equivalent, and the necessary ones will certainly be intensionally equivalent.

This is not how things seem. The four-dimensional theory of space and time does not seem to be a complicated terminological variation on three-dimensionalism (pace Miller 2006), which is what must be held if we take them both to be coherent.

Canberra Planners tend to give accounts of content that are broadly functionalist (see Braddon-Mitchell and Jackson 2006). These accounts type mental states by their typical causes as well as their typical effects (both internal effects and behavioral effects). This functional information is taken to encode propositional information. Thus the way in which a state is caused, and the kinds of effects it has on the environment when combined with other states, might type it as one that divides the worlds into those in which ice cream is present or not. But this is a coarse-grained, intensional distinction. If the Canberra Plan is to avail itself of the account of a priori metaphysics I have sketched, what will be needed is an account of content at a finer grain than the intensional which is somehow consistent with this broadly functionalist perspective.

3.1 Functionalism about Fine-grained Thoughts

How are we to make sense of this in a broadly functionalist framework? Let’s think of how functionalists individuate mental states. Functionalism individuates mental states by their input–output profiles, and how they interact with other states. More strictly it individuates them by the profiles they would have if adopted. To adopt a state is to let some piece of cognitive machinery be in charge of behavior, and interact in an unconstrained way with the rest of our cognitive apparatus. But presumably as a matter of mental reality there must be quite complex states that are not adopted, but are structured and ready to go. This is what is right about off-line simulation theory. If we can wonder whether something is right, it is because we can simulate the effects it would have on our behavior in circumstances, we can calculate what beliefs we would be committed to on equilibration between it and the rest of our current representational
states, and so on. So our mental states are states individuated by the profiles they would have if adopted.

What is the content of these states? Well, the content is got by interpretation. Functionalism as a theory of representational states, and possible worlds content, go together not automatically, but because the way in which the states affect our behavior is interpretable as fixing views of what world we are in (as when we walk through a door and reveal our view of where the gap is) and what world we would like to be in (as when we go to the grocery store so as to make our world one in which there is ice cream in the future).

Input–output profiles can, however, cut the grain of states more finely than the possible worlds. One of the ways this might be is that we count among the inputs and outputs the linguistic inputs and outputs. There can be differences in the linguistic input–output profile of a state that, under ideal interpretation, might not distinguish between worlds.

If this point is made in a crude way, it’s a way of taking a view that amounts to a blend of functionalism and the metalinguistic view. Here’s how the simple metalinguistic view might treat this case: what we need to distinguish is between whether, say,

1. libertarian freedom is possible;
2. libertarian freedom is impossible; or
3. libertarian freedom is necessary.

Suppose that libertarian freedom does exist necessarily, and that we believe this. We nonetheless must have views about what we ought to do should we come to believe one of the other options. We might think of this in one of two ways: what we are committed to should we be wrong about it being necessary, or what else we would be committed to, to maintain local consistency, should we come (mistakenly) to think that one of the other options is right.

If we are committed to the view that (3) is correct, then presumably we think that the ideal consequences of making the mistake of switching to one of the others is that eventually it would generate some kind of contradiction or some such and we would see that we were mistaken and switch back. But that does not rule out that locally we should start making certain kinds of revisions. Such revisions might be the first steps in a process of revision that will (if we are currently right and if we are ideal) lead back to where we started. But perhaps it is part of our conceptual grasp that we must begin the process in different ways, for the different possible revisions, even though in terms of possible worlds content each is just the
impossible proposition. So we begin our revisions by making judgments about the nature of freedom which would (if we were to accept the two-dimensional framework) make us have a different view as to what worlds are possible. Thus, if we tried to capture this difference in a two-dimensional table we would have to draw different tables. This is because a two-dimensional table lists all the possible worlds in its column headings and row headings—so if there are different views about what worlds are possible, then the table must have different extent. If we came to think that libertarian freedom was contingent, for example, we would draw a table with worlds that contained it, and worlds that lacked it. We might come to take the view that if it is actually instantiated then it is necessary for freedom, but otherwise not. But before, when we thought that it was impossible, we could have no worlds as actual that contained it in our table, because we took there to be no worlds that contained it simpliciter.

But what does the crude view say about these transformations? It says that the difference between (2) and (1) is the difference between coming to believe that sentence (2) expresses a truth, and coming to believe that sentence (1) expresses a truth.

The one problem with the crude view (which will be important later) is that it does not explain why coming to think that a sentence expresses a truth will make us inclined to behave in the ways summarized by the tables and produce the characteristic patterns of assertion. And of course the crude view is untenable: for we could be in a world where ‘libertarian freedom is possible’ expresses a truth. The truth could be that zebras are striped. It would be inexplicable why that changed our habits of drawing tables containing libertarian worlds. Views about isolated sentences tell us nothing important about the concept of freedom. So there is a far more general concern here. The idea is supposed to be that there is some sort of normative requirement to adopt the patterns summarized in the tables based on coming to accept (1) or (2), even on the current supposition that (3) expresses the logical truth. But what normativity would be associated with that, if these sentences do indeed express a truth, just because it turns out that they mean something else? We should not change our patterns of assertion in these ways because we come to believe something else—that the sun is yellow, that grass is green, or that Fermat’s last theorem is true. So the very conditions under which we might imagine the sentences to express a truth are the ones under which we should not change our pattern of use in the indicated way. So we had better not individuate the statements by their mere orthography, mental or written.
A more promising approach is to return to the earlier discussion of functional individuation. Suppose that any of the meaning transformations had taken place: that ‘libertarianism is contingent’ came to express a truth because, in our idiolect, it meant that zebras had stripes. Then we would expect it to have a range of consequences for our nonlinguistic and linguistic behavior. We would be puzzled at nonstriped equine things of the same size and shape as zebras. We would take there to be entailments between ‘libertarian freedom is contingent’ and claims about the equine clade.

Presumably none of these behavioral consequences is a consequence that we should expect for accepting ‘libertarian freedom is contingent’. We have a fair idea of what it would take to count as having accepted that—so we can recognize the symptomatic behavioral and linguistic consequences. And we can distinguish between the consequences of accepting ‘libertarianism is impossible’ and ‘libertarianism is contingent’ (on assumptions about the rest of someone’s mental states).

Of course it may turn out that, systematic though these differences be, they make no difference to how we should interpret the agents as taking the world to be in possible worlds terms (indeed this had better be the case). It is not as though one will walk through doors, and the other through parts of a wall that they mistakenly believe to be a door, or any such. Many differences will be linguistic and symbolic; and they will include, for those who are philosophically inclined and take two-dimensional semantics seriously, different attitudes to two-dimensional tables. So—as with any functional semantics—the normative requirement to go these different ways on the tables stems from the fact that the tables summarize our information about the recognition conditions of the mental states.

So this is why the example with zebras was too easy: the string ‘libertarianism is contingent’ can be seen to have shifted meaning to the claim involving zebras, because the effects on behavior in circumstances, and the tendency of the environment to provoke that behavior, would be of the kind that indicates that the agent takes zebras to be striped. But of course not all transformations are like that. We can functionally individuate the states even when the consequences are not ones that individuate worlds. For even supposing that we think that libertarian freedom is impossible, and thus think that the possible worlds content of ‘libertarian freedom is contingent’ and ‘libertarian freedom is necessary’ is the same—they are both the necessarily false proposition—we can recognize differences in the appropriate consequences of adopting one rather than the other, on other assumptions about mental states. Each would require that
we produce linguistic and diagrammatic output of the kind summarized by a different two-dimensional table. Of course, for those tables to be rightly interpreted as listing the possible worlds and telling us what we should call ‘libertarian freedom’ in them, we would have currently to be mistaken.

3.2 Short-term and Long-term Consequences

So the idea is that there are certain short-term consequences to appropriate behavior that characterize changing a view—even changing a view about which of two (in fact) necessarily false propositions is true. This raises a difficult issue of individuating input–output profiles here, and individuating them at different times, and it is connected with the question of idealization.

If we take ourselves to be right that libertarian freedom exists necessarily, then we must suppose that there is no idealized output profile for anyone who asserts to the contrary, for no idealized profile includes that assumption. Even if we assume that we can tinker with the ideal profile and make a small change that produces the alternative ‘libertarian freedom is contingent’, we must suppose that in the end, an ideal agent will notice that a contradiction (if these necessitations are a priori knowable) emerges and will eventually reconverge to our current state. So the ideal profile cannot differ, either because you baldly think that there is no ideal profile like that, or you think that the normatively required consequences of the change are reconvergence; so there is no way to distinguish the profile not just from other mistaken views, but from the true one!

Similarly, and perhaps more obviously, we can consider the differences it might make to our profile of behavior and assertion if we were to come to deny the four color theorem, and deny Pythagoras’ theorem. Once again, perhaps at a level of idealization and abstraction the consequences would be the same. Both denials are versions of the impossible proposition, in possible worlds semantics, and perhaps the ideal long-term consequences of revising either would be the same—syntactic contradiction in utterance, and perhaps inconsistent behaviors of other kinds that would lead ideally to reconvergence. But once again our grasp of something—and I’ll call it the concept—tells us that these two ways of being in error are individuated by the initial changes they would make to our behavior—when measuring up triangles for building, to how much ink we would order for mapmaking, to the kinds of linguistic behaviors that would engage in argument. And these initial changes are part of a dense web of changes, and they are individuated by that web.
So the fact that the ideal consequence is reconvergence does not tell us that there are not normative differences in the short-term consequences. Perhaps there’s something in what relevance logicians are going on about that is right. I can see that if I accept ‘libertarian freedom is contingent’ I should begin the process of equilibrating by being disposed to draw two-dimensional tables that contain worlds marked as libertarian, by disagreeing with those who say it is impossible, by engaging in certain disputes, and so on and so forth. Of course I may (now) confidently expect to find that beginning this way will eventually lead me into trouble, and make me revise back, but I think that nevertheless this is the way I should begin were I truly to change my mind. I must not just write down “contradiction” and revise immediately—for then I would not have changed my mind. And similarly if I revise to the view that libertarian freedom is impossible, my path to self-correction must begin in a different way. So one way in which to make sense of hyperintensional differences is via this idea of expected short-term effects.

3.3 Idealization and the Right Input Profiles
So far we have been considering the way in which unideal, short-term outputs might differ between intensionally equivalent concepts and thus allow a functionalist take on the hyperintensional. There is another way in which idealization needs to be considered. Functionalists individuate mental states inter alia by the states of the world that they are caused by. They do not, on the whole, do it by raw behavioral input. We won’t be able to narrowly specify what the right perceptual inputs are, for these will be highly context sensitive. Individuation by states of the world that do the causing is what we need when we are typing mental states by their possible worlds content. But of course real mental states in an unidealized world can only respond to (roughly) perceptually and neurally typable inputs.

This is one of the many things that leads to confusion in how to read some of the empirical literature about concepts. One example is the so-called prototype theory of concepts (which says crudely that the concept of something is a kind of mental record of the outcome of a range of sense impressions, and something falls under that concept if it falls within a kind range of allowed difference. The closer the match, the more quickly it is calculated that there is sufficient similarity). Here it often seems as though we are dealing with systems that look more like defeasible recognition systems than with concepts in any deeper sense. So, for example, the prototype of BIRD might well be something that, for most birds, gives a
reliable judgment that it is a bird. And indeed, maybe it does explain why it takes people longer to judge that a kiwi is a bird than any other thing. But, as is common ground in the literature, the system can fail. Things with good match can fail to really fall under the concept (well-made animatronic birds, parallel evolved flying foxes) and things with poor match (dinosaurs, ratlike furry avians) can fall under it. A standard thing to say about such cases is that what is required is a mixed theory (see, e.g., Laurence and Margolis 1999, chapter 1)—mixing the prototypes or whatever else it is together with some theoretical or other backup that can override. Perhaps a more perspicuous way of saying much the same thing is that what’s going on here is that we are confusing recognition systems with concepts. Of course there is an important empirical task to be done in identifying our recognition systems. But recognition systems are employed in the service of concepts, at least some of the time. Someone with expertise in eucalyptus may be quite happy to use their prototype machinery to identify eucalyptus, just as you might use a detection machine. But this does not mean that such an expert takes it to be analytic that eucalyptus are those things detected by the machine. Rather his concept EUCALYPTUS is given by something else, and he takes it to be a contingent truth that his sensory prototype is pretty good but gets it wrong if there are too many related woody myrtaceae around.

So the recognition systems, in many cases, have theoretical backups that allow us to shift from one such system to another. In some cases, the exact nature of the theoretical backups can be obscure, until such time as we have got into trouble coasting in neutral. And of course, as two-dimensionalism has taught us, sometimes the recognition systems are themselves part of the backup: as when the backup system says something like ‘what we are looking for is the natural kind, if any, that is present in the majority of the things recognized by the recognition system’. So recognition systems need not be explicitly descriptive, though in some sense presumably there must be descriptive ways of putting what it is that the systems we have in our heads are sensitive to. In some cases, of course, it may turn out that different recognition systems are in fact picking out the very same thing in every possible world. This too ought to be familiar from standard two-dimensional treatments. Where water is in fact H2O and fills the lakes and oceans, someone with the concept ‘the actual stuff that constitutes most of the sea, and can evaporate’ is tracking the very same stuff as someone who has a concept ‘the actual stuff which constitutes most of the lakes’. But the concepts are different enough, as witness what each would take to be the right thing to say about water in a world considered as actual with H2O in the lakes and
XYZ in the oceans. Of course, people who start with very different concepts in this way are very likely to revise to a common concept when it turns out that the same substance is in both places, and nothing of practical or theoretical import hangs on how they have been identified. So here we have different recognition systems that track necessarily the same thing but have different a-intensions, that is, they are terms that would have tracked different things had actuality been different.

Might we have different recognition systems that track necessarily the same thing and have the same a-intension? It seems so. The problem of hyperintensionality is perhaps just such a case. We were wrong to repudiate the idea of recognition systems as having some role in an account of concepts.

So to tie this back to what I said previously, the idea here is that recognition systems are not only perceptual. Sometimes they go by complex syntactic role. Recognition systems might go by, for example, the kind of linguistic and verbal explanations that we have, and the kind of moves we expect to be made by someone who has them. So in the case of various theoretical hypotheses about the modal status of libertarian free will we know what it takes to instantiate the concept: it takes saying various things, and being committed in various ways. We recognize users of the two distinct concepts by these features. These are going to not only fail to discriminate between worlds with respect to freedom given how things are, they are also going to fail to distinguish between worlds no matter how we suppose things to be. But what prompts us to accept that the world is one “way” or another is different. We can see that something is an argument for one view and not the other by how we respond to the inputs.

3.4 Some Final Remarks on Hyperintensionality

So here then is the picture, hazy though it be. We are looking for some kind of connection between reference, understanding, intension, and cognition. Abilities underlie the functionalist account of ordinary intensional content, and it is tempting to seek various abilities to ground hyperintensional content and thus explain how the Canberra Plan can use analysis to choose between competing claims for a priori metaphysical necessity. One candidate is the ability to say how variation in how things actually are affects how we should be disposed to label the various things that exist in all the worlds, but that is not sufficient.

The message of this section is that yes, that’s an interesting ability, but there’s another one to consider: the ability to see how claims about
necessity affect the cognitive procedures that underlie ordinary abilities that usually play a role in discrimination and cognition. Of course we have to be careful to put it that way: it shouldn’t be put as how the truth of claims about what’s necessary affects the space of worlds and how we should label them. For only one of the various two-dimensional tables we might draw given different suppositions about what is necessary or contingent actually does describe the space of worlds at all, and how we should label the things in the worlds. But each corresponds to a set of complex linguistic and cognitive dispositions, to a mental state that can, in a neutral sort of way, be adopted (and allowed to run) or not. And we can see how, if we were in the grip of one of these other than the one that we are in the grip of, we should go on.

The ability to see this is another crucial ability in the realm of the conceptual. Of course there is always some chance that the system we have adopted does not model the worlds and is not, in some sense, representational. But the idea is that sets of dispositions summarized in the tables are representation apt. Although our dispositions to two-dimensionalize are the kind of cognitive structure that is in general suitable for mapping relations between how things are and how we should label things counterfactually, it in fact does not do such a thing.

Does this mean that I want now to defend the view that the Canberra Plan should be committed to an account of concepts according to which they are dispositions to adopt different dispositions that are apt for being (but not guaranteed to be) tendencies to label things differently according to how things actually are?

No. For it seems that having told the story, that’s enough. If we work at all like this, abstract though the account is, that’s enough. There is a whole cluster of abilities here, all of which are important, and all of which need to be mentioned in giving an account of the conceptual. No one of them does all that we want from a theory of concepts.

4 Is the Analytic A Posteriori?

This account of fine-grained, hyperintensional content is one that depends on the kinds of nonideal dispositions that we might identify with various forms of concept acquisition, just as the classic intensional Canberra Plan account of concepts depends on the kinds of dispositions that we have that are idealized in a different way.

The other major problem for the Canberra Plan’s claims to be doing a priori philosophy lies in the status of these dispositions. There is a tension
here between the claims of Canberra Plan analyses discussed or offered in this volume to be a priori, and the claims that it has to be dealing in real dispositions.

In Jackson’s discussions of the Canberra Plan, much is made of reflection on our practice and of our armchair access to what we would say in circumstances. The idea is that although our concepts (not that he uses that term often: see Kingsbury and McKeown-Green in this volume) may not be explicitly introspectible in the form of definitions or even clusters, what we do have access to is what we would say in various circumstances.

Thus the idea is that though we never had the explicit idea that ‘water is the actual, local water stuff’ we were able to respond to Kripke’s Twin Earth thought experiments. Confronted with a case where some water substance exists elsewhere in the actual world (or confronted with similar counterfactual cases) we have access to the judgment that given what we mean by the term ‘water’, that other substance is not water. This thought is central to the Canberra Plan, and is part of how the apparently heartland examples of a posteriori semantics are subsumed into the a priori and analytic. It is an analytic truth\(^3\) that some substance that is not the local, actual watery stuff is not water, and this is taken to explain the very data on which the claims that the semantics of water are in every sense a posteriori are founded.

But one might think that there is an objectionable privileging of introspective access to one’s own meanings here. In the armchair one has some kind of insight, one hopes, to what one will do and judge in various circumstances. But how likely is this infallible insight? In this author’s case it certainly isn’t, and I doubt it is so in others. But in that case perhaps some of the crucial information we have about the dispositions we have to act and judge in circumstances comes from entirely a posteriori sources, like some empirical cognitive science that might give us a good guide as to what we are likely to do if actually confronted with strange cases. But if the a priori analytic is supposed to be founded on these dispositions, then its a priori status is in doubt.

There are three moves that one might make here. One is to simply insist that one’s armchair judgment is constitutive of the analytic a priori judgment. If when confronted with real events one’s judgment changes, then meaning has changed. Another is to upgrade what counts as the armchair to such a detailed envisioning of the real situation that it is hard to distinguish the two, and the armchair judgment and the real situational judgment become essentially as one. The last is to tie the analytic to the real dispositions as revealed in real situations, and give the armchair the
role of fallibly detecting these real dispositions. None of these now strikes me as satisfactory.

The first of these options requires us to hold that if actual usage comes apart from self-predicted usage, either one’s later self is in error and making a conceptual mistake, or else the concept and/or meaning has changed. Since presumably the later self in a real-life situation knows more than the armchair self, if we were to attribute error to either party, the earlier one would seem the more obvious suspect—since the error could be attributed to ignorance. Of course in the real world some kind of bias or distraction might cause error, but then this seems to be the kind of thing that requires a case-by-case analysis. As for conceptual change, once again it seems we need to know the cause of the disagreement between the armchair and the later actual self to have any idea of whether conceptual change is involved.

The second of these options makes the armchair and the real world collapse into each other, so I’ll set it aside to consider the final option, one I once mistakenly defended (Braddon-Mitchell 2004). According to this option, the armchair is a mere guide to what we are in fact disposed to do and judge, which is what is constitutive of our meanings and concepts. This seems to get right the idea that semantic properties are properties of our mental states that move us around, not merely inform our sayings in the armchair. If it turns out that when confronted with a world in which there is no single natural kind that underlies the watery stuff, I judge that it’s all water, then my armchair judgment that should I discover this I would judge that the world contained no water goes by the way. It was a theoretically infected mistake; my true concept was always there ready to guide me in the real world.

This has some plausibility going for it. Semantic properties are causally efficacious, and their most important roles are out of the armchair. There is nothing wrong with armchair deliberation—often thought experiment is all we can do. But it is, in the end, just a fallible guide to the real underlying dispositions.

There is a big “but” here, though. To take this on with full seriousness requires us to sever the link between the a priori and the analytic. For the armchair may be one place to find out what these dispositions are, but presumably some kind of brain science would in the end be a better one, able to correct for the errors of theory and predict better our linguistic and conceptual practices in real-life situations. Thus semantics becomes an a posteriori science. My meanings are not given for me to discover by introspection at all. There is a peculiar kind of analytic a posteriori take to the sorts of conceptual truths that the Canberra Plan aims to access.
I’m not sure that is now right. And it is because of the issue of how to adjudicate between a diagnosis of conceptual change and error in cases of disagreement between the earlier self and the later self. Why would we judge that there had been error attributable to either the earlier or later time slice? It could be that the later time slice takes some factors to be significant that the earlier does not, or is influenced by features of the environment that might not be relevant—such as what others do at the later time.

If these factors are insignificant, though, by whose standards are they insignificant? If the earlier time slice judges that a judgment influenced by those factors is suspect, and so does the later one, then we have a case where we can attribute error to the later time slice. But what if the later time slice has different views regarding what factors are or are not significant—perhaps the later time slice regards what others do as salient, but the earlier does not. In this case there does not seem to be any scope for attributing error, so it seems like we need to diagnose conceptual change. Thus the earlier time slice is indeed authoritative with respect to its meanings, but those will indeed change.

But notice something—if the later time slice behaves differently from how the earlier time slice predicts, there are interesting questions about how the earlier time slice should react on considering thought experiments about how the future time slice reacts. So let’s consider these matters with the example we used before. Suppose that, from the armchair, we conclude that if we were to discover that there were no underlying natural kind to the watery stuff in the actual world, we would conclude that there was no water, only watery stuff (in other words, we perhaps tacitly take water to mean ‘uniform natural kind of which most of the actual local watery stuff is composed’ or some such). In fact our armchair self believes that there is such an underlying substance, and it’s H₂O.

Now suppose that in fact, were we to discover that there were no underlying natural kind, we would conclude that water was not a natural kind, but rather a nonnatural functional kind intensionally equivalent to ‘the watery stuff’—although we judge that had there in fact been a natural kind, ‘water’ would have picked out that substance rigidly. (In other words, we perhaps tacitly take ‘water’ to be a conditional concept—something like the uniform natural kind underlying the actual watery stuff if there is one, otherwise any watery substance.)

So there appears to be a disagreement between the two time slices. On the a posteriori model mentioned above, the earlier time slice is just wrong about its concept, as revealed by its error in prediction. But there are other
armchair facts that need to be taken into account. The earlier time slice judged that, in a world in which there is no underlying natural kind to the watery stuff, there is no water. But that underspecifies such a world in salient ways. There is another possibility that can be considered from the armchair. That is the world in which there is no underlying natural kind and in which the later time slice judges that there is water. To make this a useful thought experiment, we might need to add some more factors, such as various hypotheses as to what causes the later time slice to so judge.

So here is a proposal. In a case where the earlier time slice has a judgment about the later case that is different from what will actually happen, we need to know another fact. That is what the earlier time slice would judge from the armchair about the case where the future time slice judges differently, given the causes of, and reasons for, that later judgment. Here is an alternative armchair judgment: it could, of course go either way. The earlier time slice could judge that, should the later time slice judge differently for a particular set of reasons, then the later time slice’s view should trump its own. Thus the earlier time slice may display a certain kind of deference to the later time slice, at least conditional on certain kinds of reasons and causes for the later judgment. Alternatively the earlier time slice could take the view that the later time slice is in some kind of error. If that error is not due to any ignorance on the part of the later time slice, then that amounts to the view that to make the later judgment for those reasons and causes is to have changed the subject.5

My proposal here is that the earlier armchair judgments that would be made are precisely what entitle us to think of the later judgments as being expressive of not the very same concept. So, if the later judgment (and its reasons) is one to which, when the earlier time slice considers a world as actual in which the later slice makes those judgments, the earlier slice defers, then that is why the later disposition is one which is expressive of the concept. And if the earlier slice would not defer on consideration of the world as actual where the later slice makes the different first-order judgment, then the change in judgment is expressive of a change in concept.

This brings the a priori and the analytic back in line. For the armchair judgment is not mere prediction of what the dispositions are; if the prediction is mistaken, there is the issue as to whether the disposition that will be later realized is one endorsed by the earlier time slice. And that is something about which there are a priori judgments to be had, by considering as actual worlds in which the future judgments are different from the current ones or predictions, together with the reasons and causes.
5 Final Remark

This is only a short sketch of these two sources of concern about the nature of the a priori and its connection to the Canberra Plan. But at least the field is clear for the reader to go on to consider the essays in this volume that discuss the role of a priori semantics, and a priori metaphysics as resuscitated by the philosophers in this tradition, without thinking that there are no prima facie solutions to be had to two problems which many take to be knockdown objections to even starting on the project.

Notes

1. Modulo various considerations about the definition of physical properties; definitions of a methodological kind might make it a priori that physicalism is true if it were the case that there could only be properties discovered by the methodology of natural science.

2. Where strongly empirically equivalent means that at every world where one is true so is the other, and that they are empirically equivalent at each such world.

3. At least about one meaning of ‘water’, or one concept WATER. I in fact agree with David Lewis that we have multiple such concepts, and that the word in English is ambiguous between the rigidified definite description and a purely functional term.

4. Or react in a way that would rationalize such a judgment.

5. Of course there are issues swept under the table here: one might judge that the later time slice is in semantic error, so that it in fact possesses the same concept or mean the same thing but mistakenly takes itself not to, and that this explains the difference in judgment about cases.
Dear Frank,

It’s been a while since our paths have crossed. I hope all’s well in your part of the world, and that you and your family are flourishing.

A few months back, Robert Nola contacted me to tell me that the “Canberra Plan” volume had found a publisher and to ask whether I was still interested in trying my hand at a correspondence with you—of the sort we had discussed one lovely morning on my last trip to Sydney. I said that I was and I asked whether I could do my part collaboratively with Kelby Mason. Robert said that was fine with him.

If you’re still interested in pursuing the idea, let me suggest a few vague ground rules. Our initial idea in Sydney was to exchange a few letters to try to understand where we differed about the nature of folk psychology and the relevance of various putative facts about cross-cultural variation in intuition. That’s still my interest (and Kelby’s). No doubt we’ll have plenty to debate, but I see the project less as a debate and more as an opportunity to clarify your position and mine. To that end, I suggest that Kelby and I begin by posing a few questions to which you can reply. We’ll react to your replies and respond to any questions you might raise. We can keep the process going as long as it seems fruitful. So much for our suggestions on ground rules. If you’re not comfortable with any of this, we’re certainly open to alternative proposals.

So now let’s turn to the questions. If memory serves, in the conversation which first gave rise to the idea of this exchange, you said that you did not think of folk psychology as something in the head. This took me aback, since I had always assumed that you thought of folk psychology as an
internalized theory, more or less as I did. It suggested that at least some of our disagreements about folk psychology are not real disagreements at all, since we may be talking about different things. On and off, since then, I’ve tried to come up with a charitable interpretation of your brief comments in Sydney. But I haven’t been very successful. So let me start by asking:

Q1 Do you still think that something that you’d want to call “folk psychology” is not in the head? If so, can you explain what it is?

Since it hardly seems fair to just drop these questions on you and let you do all the work, let me offer a bit of context. In a paper I wrote with Ian Ravenscroft a bit over a decade ago, we distinguished six different answers to the question: “What is folk psychology?” Four of the answers we offered maintain that folk psychology is “an internally represented ‘knowledge structure’ used by the cognitive mechanism underlying our folk psychological capacities” (Stich and Ravenscroft 1996, 132), where “our folk psychological capacities” include our capacity to attribute mental states to ourselves and others and our ability to predict and explain people’s behavior (ibid., 124–126). What distinguishes the four answers is how the information is stored. Since you maintain that folk psychology is not in the head, I assume that none of these accounts is very close to what you have in mind. Is that right?

The two remaining accounts Ravenscroft and I offered are what we called “external” accounts of folk psychology. The first of these is simply “the set of folk psychological platitudes that people readily recognize and assent to.” The second is “a theory that systematizes the folk psychological platitudes in a perspicuous way” (ibid., 132). There is a straightforward sense in which neither of these is “in the head,” which is why Ian and I called them “external accounts.” But on my (unfortunately very dim) memory of what you said when I asked for more information about your conception of folk psychology, I don’t think either of these accounts is at all close to what you had in mind either. So your conception of folk psychology is not on the chart that Ravenscroft and I constructed. Nor, I assume, do you have in mind the sort of account of folk psychology that simulation theorists urge, and that Ian and I discuss briefly in section 5 of our paper. So there are seven things you don’t mean. And at this point Kelby and I need some help. If folk psychology, for you, is not any of these, what is it?

Of course, folk psychology is only one sort of folk theory that philosophers are interested in, so our second question is simply a request for clarification:
Q2 Does what goes for folk psychology go more generally for folk theories? That is, is the story you tell about the nature of folk psychology the same sort of story you’d tell about, say, folk epistemology, folk ontology, or folk metaphysics?

If it is, then any disagreement we have here is just one instance of a broader disagreement about the nature of folk theories. So, in our discussion, we should keep in mind the broader picture, although folk psychology is probably a good test case for us to focus on.

There are two other important questions that Kelby and I would like to discuss at some point. Since your answers to them will depend on your answers to the first two questions, we don’t expect you to address them just yet. But we do want to flag them now for later discussion. On the two “external” accounts of folk psychology mentioned above, platitudes or intuitively obvious claims play a central role in characterizing folk psychology. You have also stressed the role of intuitions in conceptual analysis in other domains. And, as you know, along with a gaggle of former students, I have been exploring intergroup variation in various sorts of philosophical intuitions. The first of our two additional questions is:

Q3 What do you think would follow if there really is substantial and systematic variation in intuitions between different groups (e.g., cultural, racial, age, or gender groups)?

Our second additional question is this:

Q4 How do we decide, in the course of conceptual analysis, which intuitions are not a good guide to folk theories (and therefore should be ignored)?

The Chomskian tradition, for instance, has an answer to this question which seems well motivated, although often hard to apply: some grammatical intuitions reflect (or accord with) the rules that speakers have internalized; others are “performance errors” which result from memory limitations, failures of attention, interference from other mental systems, etc. Now, it would be hard enough to extend this Chomskian answer to folk theories in general, even if you did think that folk theories are “in the head.” Since you don’t think they’re “in the head,” it’s even harder to see what kind of alternative answer you could give. As I say, we don’t expect an answer to Q3 or Q4 yet, but we would like to return to them eventually.

OK. That should be enough to get the ball rolling. Kelby and I look forward to reading your responses.

With warmest regards,
Steve and Kelby
Dear Steve and Kelby,

Here's my belated reply to your letter of 20 February. I'll try and say as clearly as I can where I stand on folk psychology and how this connects with your questions.

1 Is Folk Psychology in the Head?

As I use the phrase ‘folk psychology’ it stands for a certain theory about what the world is like. The theory says that there are states that stand in such and such relations to each other and to events in our world. I hold that very many people believe this theory, which is why I think it is right to call it a folk theory. In the same way I think that many people hold the theory that, as a rule, unsupported bodies fall. In consequence it is right to call this a folk theory. It would not be right to call quantum physics a folk theory because very few people hold it (lots hold that it is true without knowing what it is that is true).

Is the theory that, as a rule, unsupported bodies fall ‘in the head’? Not in any natural sense. I think of theories as individuated by their contents—how things have to be for them to be true—and contents aren’t in the head—as Bob Stalnaker often remarks.

Of course, there is the question of what makes it true that some person holds some theory, and I do think that that is in the head (as a matter of contingent fact). I think, that is, that it is how a person is inside that settles what they believe, although typically what they believe concerns how things are outside them. In the case of folk psychology, I think the content—what they believe—is partly about how things are inside persons who hold the theory in the sense that the states they hold stand in such and such relations include some that they also hold are inside them, partly about the insides of other people, and partly about how surroundings interact with people's insides generally.

So what I am saying is that we need to distinguish the content of a theory from the holding of that theory when we ask if the theory is in the head. The content isn’t inside the head but what makes it the case that they hold the theory is inside the head. This applies to theories in general; it is not a thesis about folk psychology or folk theories in particular. The content of the Big Bang theory isn’t in the head but what makes it true that someone holds it is in their head. (This does not mean that the content is an intrinsic property of their head state. It is inter alia how their head state
interacts with the environment that makes it the case that they have the belief with the content that it all started with a big bang.)

Here is how this distinction affects the discussion in your letter. You mention in K&S[1] the view that folk psychology “is an internally represented ‘knowledge structure.’” From my perspective, this could not possibly be a good account of the content of folk psychology, any more than it could be a good account of the content of the folk theory that, as a rule, unsupported bodies fall. What might be true is that what makes it true that a subject believes in folk psychology is that he or she has such an internally represented knowledge structure.

Or consider one of the ‘external’ accounts of folk psychology you mention in K&S[1]: “The two remaining accounts Ravenscroft and I offered are what we called ‘external’ accounts of folk psychology. The first of these is simply ‘the set of folk psychological platitudes that people readily recognize and assent to.’” This is a view about the content of folk psychology to the effect that its content is one and the same as that of the conjunction of the platitudes as specified (not a view I hold, as it happens, but one Lewis once held, more or less). It is not in competition with various accounts of what it takes to hold the theory—that is, it is not in competition with various views about what it takes to hold a theory with the same content as the conjunction of the platitudes.

2 The Subpersonal versus Personal Level Question

On discussing the ‘in the head’ issue with Daniel Stoljar, he suggested that the question you may have in mind is the personal versus subpersonal question (and now I come to think of it, this may have been the question under discussion in Sydney).

Defenders of folk psychology often say that it is an implicit theory. I do. But what I mean by this is one, but only one, of the things sometimes meant when it is said that we have an implicit theory of grammar. There is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the nonacceptable sentences. This is the theory we make explicit by interrogating our intuitive classifications and which, when extracted and recorded in words, makes its way into grammar books as an explicit theory of grammar. That’s how grammar books get written.

What I do not mean is the sense of implicit theory in which it is said that we have an implicit theory at the subpersonal level. Our ability to classify sentences into the grammatical and the nongrammatical must have an explanation at the subpersonal level. In this regard it is like our ability
to locate sounds. The explanation for this is, in part and roughly, that our brains latch onto the relevant out-of-phase effects that occur in the inputs to our ears. I do not know if anyone knows the corresponding explanation in the grammar case but there must be one, and that is what some have in mind when they talk of our implicit theory of grammar. If we call the first the personal-level implicit theory, and the second the subpersonal-level implicit theory, what I am saying is that we have an implicit theory at the personal level and at the subpersonal level, both in the case of grammar and in the case of folk psychology.

Why do I think we have a personal-level implicit theory in the case of grammar? Because no brain science was needed to write grammar books. All the same a fair bit of work was required, so it wasn’t explicit in any obvious sense. The task was to put into finite strings of words the pattern we recognize at the personal level and that was hard work.

Why do I think that we have a personal-level implicit theory in the case of folk psychology? It isn’t explicit—if it were, how come there’s so much argument? But if all we had were an implicit theory at the subpersonal level, we would not know what we were saying when we said that someone was in pain. When I say that a sound is located at \( L \), I am not saying that the out-of-phase effects at my ears are thus and so. The content is not given by the subpersonal level account of how our brains do the locating job. Ditto for folk psychology, say I. The content is not given by the way the brain makes patterned sense of it all at the subpersonal level.

3 Does What I Say for Folk Psychology Apply to Folk Theories across the Board?
My view about folk psychology is a view about its content. The content of other folk theories will be quite different of course. So the answer is no if the question is one about similarity of content. If the question is, Do I hold that all folk theories are to the effect that various states stand in such and such relations, with the difference between folk theories being in the specified relations, the answer is also no. I think the folk hold that some things are round but that’s not a matter of relations between states.

4 What Follows if There Are Substantial and Systematic Variation in Intuitions Between Different Groups?
It would follow that there was a difference in concept (unless they were confused—more on this in later exchanges). A live example in my view is Twin Earth. There may well be subjects who insist that XYZ counts as water. In that case their concept of water differs from ours.
5 Which Intuitions Are a Good Guide?
I have no problem with the Chomskian answer you mention. I’d only add that there are two things one can mean by getting the answer wrong in the grammar case. You can mean getting the answer wrong in the sense of not conforming to the norm, or you can mean getting the answer wrong in the sense of not conforming to one’s own concept. The latter is less common than the former and it is in the latter case that performance errors most obviously enter the picture.

I hope all this helps and many apologies for being so slow.
All the best,
Frank

K&S[2]: Letter from Kelby Mason and Steve Stich, May 5, 2005

Dear Frank,
Many thanks for your letter of April 27. It did an excellent job of answering some of our questions, and raising some important new ones—all with admirable clarity and brevity. What we’ll do in this letter is (1) summarize our reading of your answer to one question (viz., “Is folk psychology in the head?”) and (2) explain why we think your comments about the personal and subpersonal levels and implicit theories raise a cluster of new questions. We suspect these new questions may underlie whatever disagreement we have on these matters.

First, however, let us note a very insignificant point that became clear to us in reading your letter. As you use the word ‘theory’, a theory can be a very small packet of information (or misinformation). For example, you write: “I think that many people hold the theory that, as a rule, unsupported bodies fall.” Other writers pack more into the notion of a theory, insisting that a theory must be a fairly complex, interrelated cluster of propositions, and some add that a theory must posit its own set of “theoretical entities.” There is, as best we can tell, exactly nothing of serious philosophical interest at stake here. We mention it only because one of us (Steve) has, in the past, been confused by comments like the one displayed above, and we suspect that other readers may have had much the same “. . . but that’s not a theory at all!” reaction.

OK. Now let’s get on to more substantive matters.

1 Is Folk Psychology in the Head?
Though we had both been puzzled, in the past, by your insistence that folk psychology [FP] is not in the head, your letter has—if we are reading
it properly—completely eliminated that puzzlement. As we read your letter, the crucial passage is: “Is the theory that, as a rule, unsupported bodies fall ‘in the head’? Not in any natural sense. I think of theories as individuated by their contents—how things have to be for them to be true—and contents aren’t in the head—as Bob Stalnaker often remarks.”

What is important, here, is that when you talk about a theory, what you have in mind is the content of the theory (or, to be a bit fussy, something that is individuated by its content), and content “ain’t in the head.” So what’s doing most of the work, here, is content externalism (aka meaning externalism), which, of course, comes in many varieties. Whether or not meaning externalism is true is a contentious matter. And one of us (Steve) has notoriously argued that the debate is deeply flawed since the participants have not said what counts as getting a theory of content right. But, fortunately, we need not debate any of these issues here. If, in claiming that FP is not in the head, you were claiming something akin to what many other philosophers have meant when they say that beliefs ain’t in the head, then we are no longer puzzled by your claim.—Progress!

2 Personal versus Subpersonal
In your letter, you suggest that “the question [we] may have in mind is the personal versus subpersonal question.” We suspect you are right about this. The distinction seems to play a crucial role in your thinking about FP and other meaning-anchoring folk theories. And, to put our cards on the table, we doubt there is a distinction to be drawn between personal and subpersonal theories (or levels, or explanations) that will do the work you need it to do.

We can think of some ways of drawing the personal/subpersonal distinction which are both reasonably clear and potentially useful for some purposes. But we don’t think the distinction drawn in these ways is the distinction you have in mind. For example, some people seem to use the “personal/subpersonal” distinction as a way of distinguishing information that is consciously (or introspectively) accessible to a person from information that is not. So, for example, if asked whether the tone she hears is coming from the left or the right, Sally will typically be able to introspect and answer with confidence. But asked whether the tone she hears in her left ear is out of phase with the tone she hears in her right ear, Sally will say that she has no idea. Introspection is no help. We don’t think that the accessible-via-introspection/not accessible-via-introspection distinction can be identified with the personal/subpersonal distinction that you have in mind, however, since, as you note in your letter, in both the grammar case
and the FP case, people can’t just introspect to determine whether a particular claim is part of their “implicit” personal-level theory.

Before going on, let us raise a pair of related questions, to be sure we’re all “still on the same page,” as they say.

**Q5** In your letter you use both ‘explicit’ and ‘implicit’ on a number of occasions. As we read you, these are, near enough, the equivalent of our ‘(readily) accessible via introspection’ and ‘not (readily) accessible via introspection’. Is that correct?

**Q6** Are we right that you do not consider the personal/subpersonal distinction to be the same as the accessible-via-introspection/not accessible-via-introspection distinction?

If the answers to Q5 and Q6 are yes, then presumably you think there are two kinds of implicit theory that might be attributed to someone: an implicit theory at the personal level and an implicit theory at the subpersonal level. However, this is the distinction that we do not understand.

To say why, let’s consider the case of grammar. Let’s assume that one of the goals of grammar is to specify the set of sentences that a speaker of a language finds acceptable from the set of sentences he does not find acceptable. (There are all sorts of quibbles that might be raised here. But perhaps they can be put to one side. “Sufficient unto the day . . .” and all that!) Let’s also agree that one main source of evidence used in trying to write a grammar for a language is (as you say in your letter) “our intuitive classifications.” So the project is to collect intuitions and construct a set of rules or principles which, taken together, entail that a (large or infinite) set of sentences are grammatical (or acceptable), and that another (large or infinite) set of sentences are not. “That,” you suggest in your letter, “is how grammar books get written.”

We’re not sure that that’s how traditional grammar books get written, since it is far from clear that traditional grammar books were intended to be descriptive rather than prescriptive. (By “traditional grammar books,” what we have in mind is the sort of books with titles like “French Grammar for University Students” that we used in university courses aimed at teaching us to speak French.) But that’s a quibble. Let’s put it to one side. There is a much more important concern about traditional grammar books.

As Chomsky argued in some of his earliest works, traditional grammar books are not fully explicit. They don’t contain enough information to enable a reader to deduce whether a specific sentence is or is not grammatical; to use them to master a language, a reader must often exploit his
or her own natural language capacity. We believe that Chomsky is clearly right to claim that traditional grammar books are not fully explicit. His solution was to try to make them fully explicit by writing generative rules and principles which would entail, for every string of words (or sounds), whether or not it was grammatical in the language. So far, we hope, there is nothing here you would disagree with. You would, we expect, agree that traditional grammars were not fully explicit, and you would agree that grammars in the generative grammar tradition aimed to be more explicit, and that (typically) they were.

But now things start to get interesting. For what was discovered as generative linguists tried to write fully explicit sets of rules and principles that would (inter alia) entail which sentences are and are not well formed in a language, is that those rules and principles had to invoke very abstract, theoretical constructs. The concepts of noun, verb, clause, etc., familiar from traditional grammar were nowhere nearly adequate (which is not to say they were not useful; they were used in some generative grammars). To be fully explicit, generative grammarians found that they had to use concepts like C-command, and X-bar, and lots of others. There is still lots of dispute among grammarians about which concepts need to be used in writing grammars of natural languages. But just about everyone in that line of work agrees that the job will require lots of very technical theoretical concepts which are quite challenging to explain and master. Indeed, one of us (Steve) used to follow that literature, but gave up because he decided that it was just too much work to try to master the increasingly difficult technical notions that were being used in contemporary generative grammars. Moreover, we believe that some of those technical concepts are so difficult to master that many perfectly competent speakers of English could not do it if they tried. Steve’s grandma was a wise and wonderful lady. But she simply did not have the intellectual wherewithal to understand C-command, any more than she had the intellectual wherewithal to understand some of the more abstract concepts used in quantum mechanics or transfinite recursion theory.

Of course it might turn out that contemporary generative grammarians have just got it wrong. Perhaps there is a way of describing the set of sentences that English speakers judge to be grammatical (or acceptable) that does not require fancy abstract concepts like C-command. But let’s assume that they are not radically mistaken in this way. That assumption allows us to ask some questions which we think are crucial in understanding your view: Suppose that GE is a “descriptively adequate” grammar of English. For every sentence, s, it entails that s is a sentence in English, or that it is
not. And these entailments match (near enough) the intuitive judgments that speakers offer about these sentences. Suppose, further, that GE makes use of technical notions like C-command, etc., which many speakers of English could not understand even if they tried to study generative grammar for a few years.

A question:

Q7 Would you say that GE is an implicit theory for speakers like Steve’s grandma? Terminology here is awkward and unsettled. If the answer is yes, would you also say that Grandma holds GE implicitly? Would you say that Grandma has tacit knowledge of GE?—We use these locutions more or less interchangeably, but perhaps you do not.

If you don’t think that GE is an implicit theory for Steve’s grandma, then we are deeply puzzled. What implicit theory did you have in mind when you wrote in FJ[1]: “There is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the nonacceptable sentences”?

If, on the other hand, you do think that GE is an implicit theory for Grandma, then other puzzles loom. For, as linguists and philosophers of linguistics have been pointing out since the dawn of generative grammar, if there is one “descriptively adequate” grammar of a natural language (as that notion was defined above) then there are many (indeed, probably infinitely many!). A grammar is analogous to an axiomatized theory (with the sentences as the analogues of theorems). And just as there are lots of ways to axiomatize a theory, so too there are lots of sets of generative rules and principles that will “generate” the same set of sentences. Let GE*, GE**, etc. be sets of generative rules and principles, each of which “captures our intuitions” about which sentences are grammatical (or acceptable) in English. (And, suppose also, that GE, GE*, and GE** are roughly equally “simple” and “elegant.”)

Q8 Would you say that GE, GE*, and GE** are all implicit theories for speakers like Steve’s grandma?

We’re not making any bets on how you would answer this. But, as you know, in the Chomskian tradition (which was a major influence on the emergence of contemporary cognitive science), linguists did not rest content with the idea that GE, GE*, and GE** were all implicit theories for Grandma (or, as they preferred to say, they were not all “tacitly known” by Grandma). They insisted that (at most) one of these was the right grammar, even though all of them “captured the intuitions.” Which one
is it? Their answer was that it is the grammar that is actually represented in the speaker's mind. There's lots of theoretical baggage in the background here: The mind is (a bit) like a computer. There are rules that really are stored in the mind (just as there are rules that are really stored in a computer). And those rules play a causal role in producing the intuitions. If GE is the right grammar, and GE* etc. are not, that is because (a representation of) GE really is stored in the speakers' minds and really does play a causal role in producing intuitions (and comprehension, etc.), while GE*, etc. do not. The passage we just quoted from FJ[1] suggests that you might have some sympathy with this idea. For you talk about an “implicit theory that drives our classifications” and “drives” certainly invites a causal interpretation. So it is time for another question:

**Q9** Do you think that the right implicit theory to attribute to Grandma is the one that is really represented in her mind, and which plays a causal role in generating intuitions?

If the answers to Q8 and Q9 are both no, then we're flummoxed. What do you mean when you attribute an implicit theory to Grandma? But if the answer to Q8 is no and the answer to Q9 is yes, still more puzzles await. For if the right implicit theory to attribute to Grandma is the one that is really represented in her mind, then there is good reason to suppose that merely probing Grandma's intuitions will not be enough to tell us which one it is. That's useful data, to be sure. But it is not likely to help much in deciding between GE, GE*, GE**, etc. To do that, linguists and cognitive scientists have insisted, we have to look at lots of other sorts of data. Chomsky, famously, thought we have to look at other languages, in an effort to find linguistic universals. (One of us—Steve—used to make his living arguing that this strategy was not likely to succeed, though he has since come to have doubts about his earlier view.) Other linguists and psycholinguists have proposed other strategies, including looking at developmental data and at data from people with impaired language capacities, in an effort to see what the rule system looks like when it is broken or not fully in place. Still others have developed various experimental techniques, using reaction times, Stroop effects, and a whole bunch of other tricks. More recently, some have begun to use brain imaging data in an effort to figure out which rules are actually represented in the mind. No one doubts that it will be hard work to figure out what rules and principles are actually causally responsible for intuitive judgments about grammaticality (etc.). But it seems clear (to us, at least), that once one makes the "realist" move which maintains that the right grammar to attribute to Grandma is the one which is actually represented in her mind, then all sorts of evidence
become potentially relevant. If one is a “realist” about grammar, determining the correct grammar for a language (or a speaker) is not something that can be done from the armchair.

Finally, let’s return to the distinction between personal and subpersonal levels which we find so puzzling. Suppose, again, that your answer to Q9 is yes and further suppose, for sake of argument, that GE is the implicit theory really represented in Grandma, and the theory that plays a causal role in generating intuitions. Our final question is this:

**Q10** Would you call GE a personal-level theory, or a subpersonal-level theory?

We suspect that you would have to place GE at the personal level, based on the passage we’ve already quoted from FJ[1]. There the context indicates that you’re talking about personal-level theories, so it seems that what drives our classifications is the implicit personal-level theory. If Chomsky’s right, then the theories codified in traditional grammar texts are far too impoverished to drive our classifications, whereas, ex hypothesi, GE does drive our classifications. Thus it seems that, on your account, GE is a personal-level (implicit) theory.

But then we don’t know what’s going to be left at the subpersonal level. It can’t be that subpersonal theories are those which aren’t readily accessible to consciousness, since you do seem to allow personal-level theories which aren’t readily accessible to consciousness either (i.e., implicit personal theories). Nor can it be that the subpersonal theories are those which refer to properties unfamiliar to Grandma. For C-command structures are unfamiliar to Grandma, and we are assuming that you would nonetheless place GE at the personal level. Even if you don’t spot us this assumption, you do allow that personal-level theories might be implicit, and discovering them might be “hard work,” which suggests that they might involve unfamiliar properties. So what could be left at the subpersonal level?

We believe that just about all of what we’ve said about grammar could be said, mutatis mutandis, about folk psychology. But that’s best left for another letter. This one has gone on long enough.

All the best,
Kelby and Steve

---

**FJ[2]: Letter from Frank Jackson [2], May 7, 2005**

Dear Kelby and Steve,
Your letter of 5 May helps a lot—many thanks. I see that there are a number of key terms that get used differently by different players and that’s been
the source of some of the disagreement and a fair bit of confusion. This
will be a letter mainly about how I’m using the terms. I’ll set my comments
against some of your questions.

In K&S[2] you say: “As you use the word ‘theory,’ a theory can be a very
small packet of information (or misinformation).” Yes. I was moved many
years ago by Quine’s decision to use ‘object’ very inclusively and when it
was important to be exclusive, to use ‘unified object’ or ‘object of interest’
or. . . . I follow the same policy with ‘theory’. You are right that many don’t
follow this policy and use ‘theory’ for something that is by definition
complex and posits entities that play various roles. But I note that many
also say things like “Folk psychology is a complex theory that posits . . .”
as if the words ‘complex . . . that posits . . .’ added something. I agree that
it would be good to be explicit about the usage question to avoid needless
confusion.

Also in K&S[2] you say: “If, in claiming that FP is not in the head, you
were claiming something akin to what many other philosophers have
meant when they say that beliefs ain’t in the head, then we are no longer
puzzled by your claim.—Progress!” Yes I think we have progress but let me
sound a word of caution. When people say that beliefs, in the sense of
belief contents, ain’t in the head, they can have two different things in
mind (well, many, but two relevant here). One is that content per se ain’t
in the head. Belief content is how things are being represented to be and
that’s not in the head. This is what I had in mind when I mentioned Bob
Stalnaker. The other is that belief content need not be shared between
doppelgangers in our world (not that there are any of course). I agree with
the first claim but deny the second. Many hold that the first entails the
second—Bob Stalnaker is an example I think—so the difference matters
little for them but it matters for me. How can I deny the clear message of
Twin Earth etc.? I’d send you the key papers except that we’ve enough on
hand.

You raise the question in K&S[2]: “Q5 [Are ‘explicit’ and ‘implicit’] the
equivalent of ‘(readily) accessible via introspection’ and ‘not (readily)
accessible via introspection?’” No; but let me answer the next question
before I explain what I mean: “Q6 Are we right that you do not consider
the personal/subpersonal distinction to be the same as the accessible-via-
introspection/not accessible-via-introspection distinction?” Yes, but the
bald answer may confuse. Let me first say what I mean by the personal/
subpersonal distinction. I’ll use the sound location example.

When someone hears a sound as coming from location L and accepts
that things are as their perceptual experience represents them to be, then
they believe the sound is at $L$. I say that is representation at the personal level simply to flag that it is the content of what they believe. There’s nothing more to the term ‘personal’ than that. There is a connection with introspection in that people are often rather good at accessing what they believe. But in cases where they do not know what they believe—self-deception, etc.—I’d still say that their belief that $p$ was personal level even if they had no idea that they believed that $p$. (I doubt if anything important for our discussion hangs on this last point.)

When someone hears a sound as coming from location $L$, their brains carry the putative information that the sound comes from $L$ by virtue of carrying out-of-phase information $PI$ indicative of location $L$. I say this is subpersonal information and representation at the subpersonal level because (1) they do not believe that $PI$ (nor do they perceptually represent that $PI$), and (2) the information is stored in the brain. The connection with introspection is simply that subjects’ failure to access $PI$ without fancy brain science is good evidence that they do not believe that $PI$. I suppose that it might be argued that they do believe that $PI$ and this is one of the cases where belief content is inaccessible (ditto for perceptual representation), but this would seem a very strained position.

Now for what I mean by the implicit/explicit distinction. It is all to do with the availability to the subject of sentences that capture what they believe—of, that is, sentences that represent as their minds do when they believe $P$, where $P$ is the theory we are talking about. Of course some philosophers have a highly linguistic conception of belief that means that there must always be a suitable sentence available but I belong to the party that thinks that dogs have beliefs and that we have many beliefs that outrun our linguistic capacities—obvious examples being our perceptual beliefs about color shades.

Now for the grammar example. When presented with sentences of English: $S_1$, $S_2$, . . . I am able to classify them as grammatical or the opposite. Let’s pretend I am infallible to avoid irrelevant complications. What do I believe when I believe that $S_{17}$ is grammatical? Three answers might be offered. (1) I believe that $S_{17}$ is one of the sentences that is tagged ‘grammatical’ by competent users of English. (2) I believe that $S_{17}$ plays the “OK” role in its language. (3) I believe that $S_{17}$ exemplifies the pattern (disjunctive pattern) that unifies the “OK” sentences in English. One way to highlight the difference between (2) and (3) is to note that playing the OK role in its language is shared between grammatical sentences in English and grammatical sentences in Japanese, but the pattern is not.
We know there must be a pattern that unifies the grammatical sentences in English, otherwise we could not acquire the ability to recognize the grammatical sentences in English after a finite number of presentations. Ditto for Russian etc. (Linguists have of course many interesting things to say about the paucity of the evidence base and what that does or does not tell us about the brain, evolution, etc.)

Now consider my belief that S17 is grammatical in sense (3), the sense in which I believe that S17 falls under a certain complex pattern that I can recognize. Do I know what the pattern is? In one sense no. I cannot give an open sentence ‘X is . . .’ which represents that X has the pattern (or cannot give a set of open sentences, but let’s think in terms of a single complex open sentence) that captures its nature. All I can do is produce is ‘X is grammatical’. On the other hand I know case by case what makes a sentence fall under, or fail to fall under, the pattern. I am not in the position of someone who says “I know ‘She are happy’ is crook but search me how to fix it.” I can say case by case for each crook sentence how to fix it (there will be many ways of course) and that case-by-case information is enough to construct in principle the sentential representation. That’s what I mean when I say that I have an implicit knowledge of grammar. My situation with respect to grammar is different from the situation of chicken sexers (assuming that the story philosophers tell about chicken sexing is correct).

Now some comments that bear on your good questions.

(1) The case is not one where I believe but am unaware that I believe. I believe that S17 is grammatical (in sense 3), and I believe that I so believe. What I cannot do is produce an open sentence—other than ‘X is grammatical’—that gives the pattern.

(2) What’s implicitly known is the pattern, not the open sentence. What’s implicitly known is what it is to be grammatical, and that’s the pattern and not the sentence. This is why I do not discuss the very interesting question as to whether or not some open sentence that does the job contains terms expressing concepts that I do not possess.

Consider a child who can recognize circles and can do more than merely recognize them. For each plane figure they can say whether or not it is a circle and why—“It’s that bump that rules it out.” If you draw plane figures on infinitely stretchable rubber, they can distort the rubber to make them circles, etc. What they cannot do is produce ‘X is a circle iff X is a plane closed figure with the maximum area to perimeter value’, or ‘X is a circle iff X is a plane closed figure with such and such a tangent property’, or ‘X is a circle iff X is a plane closed figure whose perimeter is everywhere equidistant from a single given point’, etc. This child will have an implicit
grasp of what a circle is by my lights but may well not have the concept of, say, a tangent.

(3) When I say that “[t]here is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the nonacceptable sentences,” I mean that it is the pattern that drives the classifications—not some set of sentences. Of course there will be two patterns—the one in the sentences and the one in the brain. This follows from the fact that the classifications go via the brain. I am talking about the one in the sentences.

There is an interesting question about the pattern in the brain that can be raised in terms of the possibility (the one you mention, and as you say it is more than a possibility) that there are a number of open sentences deploying different concepts that are equally good at picking out the pattern that is being grammatical. I’ll raise it in terms of the circle example, as the same applies in that case.

There will be a story about how the child’s brain stores the information that some closed figure is a circle, and the different open sentences of the form ‘X is a circle iff . . .’ may well differ in how closely they mirror how the brain does the job. Perhaps the brain latches onto the tangent property, or the area property, or the equidistant from a given point property, . . . These are the kind of issues David Marr talks about of course.

(4) One might say that if there are different open sentences that capture the pattern that is being grammatical, there are different patterns in nature that are equally candidates to be the pattern I have an implicit grasp of. But consider the circle again. There are not different shapes for each open sentence. I say the same in the grammar case.

(5) One might point out correctly that often people are wrong about what drives classifications. But I do know what makes me say ‘She are happy’ is not grammatical. As I say above, the case is not like the chicken sexing case.

Best to you both and hope this helps—pity we can’t do this “across a table” but email helps.

Yours,

Frank

K&S[3]:  Letter from Kelby Mason and Steve Stich [3], June 18, 2005

Dear Frank,

One of our motives for initiating this correspondence was the suspicion that some of the apparent disagreements between us on issues in the philosophy of mind might be traceable to the fact that we use some crucial
terms, like ‘folk psychology’ and ‘tacit (or implicit) theory’, in very different ways. Your last letter has reinforced this suspicion. But it has also raised lots of questions about how, exactly, you do use these terms. We’ll devote this letter to setting out these questions.

The bulk of your last letter was devoted to explaining what you mean by two distinctions: the personal/subpersonal distinction and the implicit/explicit distinction. By and large, we found your account of the first of these distinctions to be unproblematic (with a caveat or two to be noted). By contrast, we found your account of the second distinction to be deeply perplexing. We’ll start with the personal/subpersonal distinction and then go on to the implicit/explicit distinction.

1 Personal/Subpersonal

In FJ[2] you say:

When someone hears a sound as coming from location \( L \) and accepts that things are as their perceptual experience represents them to be, then they believe the sound is at \( L \). I say that is representation at the personal level simply to flag that it is the content of what they believe. There’s nothing more to the term ‘personal’ than that. There is a connection with introspection in that people are often rather good at accessing what they believe. But in cases where they do not know what they believe—self-deception, etc.—I’d still say that their belief that \( p \) was personal level even if they had no idea that they believed that \( p \). (I doubt if anything important for our discussion hangs on this last point.) (FJ[2], emphasis added)

So, if we understand you correctly, a personal-level \( X \) (for a given person, \( P \)) is simply an \( X \) that is the content of (one or more of) \( P \)'s beliefs. One question that arises here is: What kinds of things can be personal-level \( X \)s? Since a personal level \( X \) must be a content of a belief, we assume that the answer is that all personal-level \( X \)s are representations or contents—when these are understood, as explained in FJ[1], as things that are not in the head. Since you identify a theory with its contents, you can also talk about a personal-level theory. (Indeed, given your broad use of ‘theory’ every personal-level \( X \) is a personal-level theory.)

Before turning to the more problematic implicit/explicit distinction, we want to raise three concerns about your account of the personal/subpersonal distinction. The first of them will probably have no bearing at all on the sorts of issues we’ll be discussing in this correspondence; we raise it because it might be relevant to understanding some of your views on other matters. If the notion of the personal level is tied to belief in the way we’ve stated above, then there can’t be personal-level representations that are not propositional (or, if you prefer a less jargon-laden way of making the point, there
can’t be personal-level representations that are not possible contents of beliefs). So, for example, since one can’t believe a concept, there are no personal-level concepts. And since one can’t believe the content of a taste sensation, the contents of taste sensations can’t be personal-level things either.

Is this really the way you want to use the notion of the personal level? We rather suspect that it isn’t, and that your explanation of the personal level, which appeals to the contents of what is believed, was intended only for the special case of propositional representations or contents. That’s fair enough, though it does leave you with the job of giving an account of the distinction between personal-level and subpersonal-level representations when those representations are not propositional.

A second concern, closely related to the previous one, is that the interpretation we’ve proposed of the passage from FJ[2] is almost certainly too literal minded in restricting itself to belief, since your personal/subpersonal distinction would presumably hold for attitudes like desires as well. For instance, suppose Waugh wants to throw the ball at the stumps, in order to run out Muralitharan. One representation in Waugh’s mind is “I [Waugh] throw the ball at the stumps.” At the same time, his mind will also contain some representation with more motoric detail, along the lines of “I throw the ball with a velocity of at least \( x \), so that it traces a trajectory of \( y \) . . .” We suspect that you would call the first representation personal and the second subpersonal. So a more charitable interpretation of the passage we quoted would be: a personal-level \( X \) is the content of one or more of \( P \)’s beliefs or desires (or perhaps his propositional attitudes in general?). We raise this concern just to make sure we’ve understood your distinction between the personal and subpersonal levels; in this letter we’ll continue to focus on belief as the attitude we’re all interested in.

Our third concern is that tying the notion of the personal level to the notion of belief could lead to some question-begging mischief down the road. Why? Well, one of the reasons we are all interested in getting clear about the notions of implicit theory and folk psychology is that these notions play a central role in the debates over eliminativism. And eliminativists, of course, maintain that there are no beliefs. If they’re right, then on your account of the personal level, there are no personal-level representations either. So if one is going to make claims about personal-level representations in laying the groundwork for the eliminativism debate—in saying what folk psychology is, perhaps—then one has to be very careful not to beg the question. The devil is in the details, of course. Whether a question is indeed being begged will turn on the details of the argument at hand. All we want to do here is to raise a caution flag.
2 Implicit/Explicit

We turn now to your account of the implicit/explicit distinction, which, as we mentioned earlier, we find very perplexing. To explain why we’re perplexed, we’ll raise four (clusters of) problems, going from less to more serious.

Q11 To what sorts of things does the distinction apply?

As in the case of the personal/subpersonal distinction, we are not entirely clear about the sorts of things to which you would apply the terms ‘implicit’ and ‘explicit’. The clear case, of course, is theories. You say: “there is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the nonacceptable sentences” (FJ[2], emphasis added).

And elsewhere you make it clear that you think folk psychology is also an implicit theory. So theories, which, for you, are propositional (?) representations or contents, or clusters of such representations, are one kind of thing that can be implicit or explicit. But in FJ[2] you also talk, on several occasions, of implicit knowledge of a pattern. (“What’s implicitly known is the pattern . . .”; “there are different patterns in nature that are equally candidates to be the pattern I have an implicit grasp of.”) Later in this section, we’ll raise some concerns about your notion of a pattern. But whatever a pattern is, it seems unlikely that patterns are theories. (One can’t believe a pattern; it seems odd to say that a pattern is true or false.) So there are at least two kinds of things to which you would apply the terms ‘implicit’ and ‘explicit’. Are these the only two? Or are there more?

Q12 What is “availability”?

Here is the passage from FJ[2] in which you give your most explicit account of the implicit/explicit distinction:

Now for what I mean by the implicit/explicit distinction. It is all to do with the availability to the subject of sentences that capture what they believe—of, that is, sentences that represent as their minds do when they believe that P, where P is the theory we are talking about. Of course some philosophers have a highly linguistic conception of belief that means that there must always be a suitable sentence available but I belong to the party that thinks that dogs have beliefs and that we have many beliefs that outrun our linguistic capacities—obvious examples being our perceptual beliefs about color shades.

We’ll shortly raise some questions about just what kind of sentence has to be available for a theory (or pattern) to be explicit. But first we want to ask what you mean by ‘availability’. There are three ways to unpack the idea
that there are sentences available to a subject which capture what he believes. In the strong sense, a subject has such sentences available when he can produce them if asked (with, perhaps, some idealization about full rational reflection and the like). In the moderate sense, the subject need only be able to recognize them as sentences which do indeed capture what he believes. In the weak sense, all that’s needed is that there be sentences in the subject’s language which will do the job, even if he couldn’t produce them when asked, or recognize them as being the right sentences.

We assume that when you talk of sentences being available, you mean available in the strong sense. The weak sense is obviously too weak, since it would make too many theories explicit. The moderate sense also seems too weak. To see why, take your example of the child who can tell circles from noncircles but who has no beliefs of the form ‘X is a circle iff X is a plane closed figure with the maximum area to perimeter value’. Let’s assume that the child—call him Meno—has had some mathematical education, so he understands all the key terms in this sentence—‘plane closed figure’, ‘maximum’, ‘area’, ‘perimeter’. Suppose now someone tells Meno the sentence just quoted, viz. “X is a circle iff . . .” and, after quick reflection, Meno agrees. In the moderate sense of ‘available’, this sentence was available to Meno and thus Meno had an explicit theory of circles. Since you don’t think someone like Meno (before his enlightenment) has an explicit theory of circles, we infer that the moderate sense of ‘available’ is also too weak.

Though we think that the strong sense of ‘available’ is what you intended, your examples of dogs and perceptual beliefs convinced us that perhaps we had best ask whether this is right. For in both these cases the required sentences are not available in the strong (or moderate) sense because they are not available in the weak sense. So, before going on, we want to be sure we understand you. When you talk of appropriate sentences being available, do you, as we assume, interpret ‘available’ in the strong sense, viz. there not only have to be suitable sentences in the subject’s language, but the subject has to be able to produce them when asked to produce a sentence which captures what he believes?

Q13 What kind of sentences have to be available?

OK. Now that those preliminary concerns are on the table, let’s turn to what we find most problematic about your account of the implicit/explicit distinction. To begin, let’s quote, again, the crucial passage from FJ[2]: “Now for what I mean by the implicit/explicit distinction. It is all to do with the availability to the subject of sentences that capture what they believe—of, that is, sentences that represent as their minds do when they believe that P, where P is
the theory we are talking about." So if someone can produce sentences which "capture" what he believes, then his theory is explicit, while if he cannot, then his theory is implicit. And to "capture" what he believes, the sentences have to "represent as their minds do when they believe that \( P \), where \( P \) is the theory we are talking about." Right after this passage, you apply the idea to the example of grammar. And it is there that we get well and truly lost. Here is part of what you say: "Now consider my belief that S17 is grammatical in sense (3), the sense in which I believe that S17 falls under a certain complex pattern I can recognize. Do I know what the pattern is? In one sense no. I cannot give an open sentence ‘\( X \) is ...’ which represents that \( X \) has the pattern (or cannot give a set of open sentences, but let’s think in terms of a single complex open sentence) that captures its nature" (FJ[2]).

Now one might think that it is trivial to give the sort of open sentence you require, since surely ‘\( X \) is grammatical’ is an open sentence which represents that “\( X \) has the pattern ... that captures its nature.” But clearly that is not what is required, since your very next sentence is: “All I can do is produce is ‘\( X \) is grammatical.’” But if ‘\( X \) is grammatical’ won’t do, why not? Surely it does “represent that \( X \) has the pattern that captures its nature.” So we infer that there must be some additional constraints on the open sentence that must be available. But what could those constraints be? Here it would be easy enough to generate a long list of uncharitable proposals, and to show that they won’t work. Perhaps the most obvious candidate is the constraint that the sentence can’t use the word ‘grammatical’. OK. Then you can say ‘\( X \) is a sentence and \( X \) is not ungrammatical’. Or ‘\( X \) is a well-formed sentence in my idiolect’. Or ‘\( X \) is a sentence in the dialect of English that I speak’. Or ‘\( X \) is a sentence generated by the internally represented set of linguistic rules that plays a central role in my language processing’. Or ... But there is really no point in offering uncharitable proposals and showing that they won’t work. Our goal isn’t to argue against your view but to understand it. And here, we must confess, we are just flummoxed. So let us try to pose the question that puzzles us simply and clearly as possible:

**Q14** What are the constraints on the “belief capturing” open sentence that must be available to a person if his theory is to count as explicit rather than implicit?

Before leaving this section, it might be useful to offer a very rough sketch of the account of explicitness that, we believe, plays an important role in linguistics and psycholinguistics. As we noted in K&S[2], early on in the history of generative grammar, Chomsky argued that traditional grammar books were not fully explicit; to determine what they claim about a specific sentence (or
sequence of words) in the target language, one often has to make sophisticated use of one’s own linguistic capacity. To make grammars fully explicit, Chomsky proposed that we try to discover a set of generative rules and principles which would entail, for every string of words (or phonemes), whether or not it was well formed in the language. The notion of entailment invoked here was closely related to the notion of entailment assumed in math and computer science. Roughly speaking, for every grammatical sentence in the language, there has to be a formal proof from the rules and principles to the claim that ‘S is grammatical’. And, even more roughly speaking, the proof is formal in the sense that its correctness can be checked by a computer. Back in the old days when Steve was young, a grammar that was fully explicit and that correctly specified the set of sentences in the target language was said to be “descriptively adequate.” But, as we noted in K&S[2], for a variety of reasons people in the Chomskian tradition were not satisfied with the goal of producing grammars that were merely descriptively adequate. Some of them, including Chomsky, insisted that we should aim at discovering the grammar that is actually represented in the minds of speakers of the language and which plays a causal role in producing linguistic intuitions, and in more important jobs like comprehension and speech production.

Now in the quote from FJ[2], you say: “It is all to do with the availability to the subject of sentences that capture what they believe—of, that is, sentences that represent as their minds do when they believe that P, where P is the theory we are talking about.” And taken out of context, the passage we have emphasized might be read as claiming that what must be available is what the Chomskians want—viz. a statement of the rules that are actually represented in the speaker’s head. But in light of everything else you say, we take this interpretation to be unlikely. Another interpretation of your view on what must be available for a theory to be explicit, is that what is required is something like what the Chomskians call a descriptively adequate grammar. That may be closer, but we doubt you’ll embrace that proposal either. So, rather than speculating further, let us bounce it back to you. How should your account of the explicit/implicit distinction be interpreted? What is the answer to question Q14?

Q15 What are “patterns”?

So far, we have said very little about “patterns,” though they loom large in your discussion of the implicit/explicit distinction. Here, again, we find what you say quite puzzling. We’ll begin by noting what might just be a slip on your part, though it might reflect some much deeper point that we haven’t understood. In discussing implicit knowledge of grammar, in FJ[2],
you say: “I know case by case what makes a sentence fall under, or fail to fall under, the pattern. I am not in the position of someone who says ‘I know “She are happy” is crook but search me how to fix it.’ I can say case by case for each crook sentence how to fix it (there will be many ways of course) and that case-by-case information is enough to construct in principle the sentential representation.”

Now it is surely true that, presented with some cases of “crook” sentences, you know how to fix them. But, if “crook” means ungrammatical, then we think it is simply false that, for each crook sentence, you know how to fix it. Every sentence in Hungarian is an ungrammatical sentence in English. But presented with a sentence in Hungarian and asked to “fix it,” we would be stymied. We would have no idea what we were supposed to do and, we assume, you wouldn’t either. Perhaps you want to restrict your claim to “crook sentences” that are made up of English words. But even here, we think your claim is false. It is easy enough to “fix” your example, ‘She are happy’. But how would you “fix” something like ‘The the the in loud a a’? Like you, we judge this sequence of words to be ungrammatical. But asked to fix it, we’d be no better off than we would be in the case of the Hungarian sentence.

A bit later, you make some related remarks about “a child who can recognize circles and can do more than merely recognize them. For any plane figure they can say whether or not it is a circle and why—‘It’s that bump that rules it out’” (FJ[2], emphasis added). Here again, we’re puzzled. There are no doubt lots of kids who can do just what you say for figures like this one:

![Diagram of a circle and a figure]

But what is the kid supposed to say when asked why the following figure is not a circle?
If we were asked this question, we would not have a clue about how to answer (unless: “Because it isn’t round!” counts as an answer).

Is any of this of any importance? Well, frankly we’re not sure. What is clear is that the examples of the “crook sentence” and the figure that is not a circle loom large in your discussion of patterns. Moreover, patterns play a central role in your account of what is implicitly known in the case of grammar: “(2) What’s implicitly known is the pattern, not the open sentence. What’s implicitly known is what it is to be grammatical, and that’s the pattern and not the sentence” (FJ[2]). And you seem to think that the grammar case and the circle case are deeply analogous.

We find all of this deeply puzzling. To explain why, let’s focus on the case of grammar. In one example in FJ[2] you say: “I believe that S17 exemplifies the pattern (disjunctive pattern) that unifies the OK sentences in English.” And a bit later, you elaborate on this as follows: “We know there must be a pattern that unifies the grammatical sentences in English, otherwise we could not acquire the ability to recognize the grammatical sentences in English after a finite number of presentations. Ditto for Russian etc.”

But what could it mean to say that there is a pattern that unifies all the grammatical sentence of English? Well, one thing that is surely true is that there is a set of sentences (or phoneme sequences) which are grammatical in English. Another undisputed truth is that there infinitely many sentences in that set. And it is widely held that considerations of learnability like those you cite indicate that the set of grammatical sentences must be recursive (or, a bit more modestly, that it must be possible to characterize them with a finite set of rules or principles). Finally, as we noted in our previous letter (K&S[2]) if there is one finite set of rules that generates a given infinite set, there are indefinitely many. As far as we can see, these are all the facts there are in this vicinity (as Dave Chalmers likes to say). So when you talk about a pattern that “unifies” the grammatical sentences in English, which of these facts do you have in mind? Is it just a way of saying that the sentences of English are members of an infinite recursive set (or an infinite set that can be characterized by a finite set of rules)? If that’s not what you mean, then we need some help, since we can’t think of anything else that you might mean. But if that is what you mean, then problems loom elsewhere.

For consider (2) above. You say that what is implicitly known is the pattern. On the reading we’re proposing, what that means is (something like) what is implicitly known is an infinite set that can be characterized by a finite set of rules. So far, so good. Although we should note in passing
that ‘implicit knowledge’ in this context strikes us as a bit odd, since to say that a person implicitly knows the set (= the pattern) in this context is simply to say that she can (under idealized circumstances) correctly judge that each sentence in the set (and only these) is grammatical. And that is simply an observation about how the person behaves (under idealized conditions). One might have thought that positing implicit knowledge was a way of explaining a range of behaviors; but on this reading, talk about implicit knowledge does no serious explanatory work at all. It is just a way of describing the behavior that the subject would exhibit.

That this is more than just a terminological curiosity emerges in the following quote: “(3) When I say that ‘[t]here is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the non-acceptable sentences,’ I mean that it is the pattern that drives the classifications—not any set of sentences” (FJ[2]). (The context suggests that in the last bit, “not any set of sentences,” you’re talking about the “open sentences” discussed earlier, which, are not “available” to subjects in cases of implicit knowledge. Though nothing we say will turn on this reading.)

Now, as we noted in K&S[2], ‘drives’ invites a causal interpretation. And here we are completely stumped. How could it be true that an infinite recursive set (or an infinite set that can be characterized by a finite set of rules) drives (i.e. causes) our classification of sentences? What does the driving (at least if anything like the Chomskian story is correct) is the specific set of recursive rules that is represented in the speaker’s brain. The sentence that follows (3) leaves us even more confused: “There is an interesting question about the pattern in the brain that can be raised in terms of the possibility (the one you mention, and as you say it is more than a possibility) that there are a number of open sentences deploying different concepts that are equally good at picking out the pattern that is being grammatical” (FJ[2], emphasis added). Here again, the reading of your “pattern”-talk that we proposed above makes no sense. If a pattern is an infinite recursive set, then we’re reasonably confident that there are no patterns in our brains since our brains are distressingly finite.

3 Conclusion

When we started to write this letter, we planned to end with a sketch of the account of “tacit” theories that we favor, and compare it to your account of “implicit” theories. But for two reasons, we now think that is best saved for another occasion. First, we don’t really understand your account well enough to do a compare and contrast. Second, this letter is already very long. So we’ll close with the hope that your response will
dissolve some of the puzzlement we’ve expressed about your account of
the explicit/implicit distinction.
All the best,
Kelby and Steve

FJ[3]: Letter from Frank Jackson [3], June 21, 2005

Dear Kelby and Steve,
I’m replying quickly to your letter of 18 June in the hope of heading off some misunderstandings.

1 Personal versus Subpersonal
I illustrated this distinction with the case of belief because that is what we’ve been discussing. I don’t think that personal-level states are one and all beliefs. I think being a belief is sufficient for being a personal-level representational state but not necessary—and I take it this is standard doctrine. I think, for example, that perception is a personal-level representational state, and one can perceptually represent that $p$ without believing that $p$.

It may also help if I comment briefly on “your three concerns about your [my] account of the personal/subpersonal distinction”: “If the notion of the personal level is tied to belief in the way we’ve stated above, then there can’t be personal-level representations that are not propositional (or, if you prefer a less jargon-laden way of making the point, there can’t be personal-level representations that are not possible contents of beliefs). So, for example, since one can’t believe a concept, there are no personal-level concepts. And since one can’t believe the content of a taste sensation, the contents of taste sensations can’t be personal-level things either” (K&S[3]).

I distinguish representational states from representations. I think all representational states are propositional (although not because I believe all representational states are beliefs). I don’t think all representations are propositional. I doubt if there’s any substantial difference between us here. It is a question of terminology. By the way I think you can believe the content of a taste sensation. When something tastes sweet to me I may well believe that it is sweet: “your personal/subpersonal distinction would presumably hold for attitudes like desires as well” (ibid.). Yes. Again: “Our third concern is that tying the notion of the personal level to the notion of belief could lead to some question-begging mischief down the road” (ibid.). I agree. The personal versus subpersonal distinction would apply even if certain eliminativist views turned out to be correct.
2 Implicit/Explicit

(i) You ask in your previous letter: “To what sorts of things does the distinction apply?” and worry about my talk “of implicit knowledge of a pattern . . . [and say] . . . it seems unlikely that patterns are theories. (One can’t believe a pattern; it seems odd to say that a pattern is true or false.) So there are at least two kinds of things to which you would apply the terms ‘implicit’ and ‘explicit’. Are these the only two? Or are there more?”

I think you’ve been foxed by my phrasing. The distinction applies to theories (in my inclusive use of that term). Knowledge of a pattern is knowledge about where and when the pattern is exemplified and this is a theory and can be true or false (and will be true if it is indeed a case of knowledge). Think of belief as to where the party is. You might say that one cannot believe where a party is, and that it seems odd to say that where a party is is true or false. But of course belief as to where a party is is any belief of the form ‘the party is at $x$’ and will be true (false) just if the party is at $x$ (is not at $x$). Likewise, when I talk of knowledge about a pattern, I mean knowledge to the effect that $x$ falls under the pattern, and it will be true or false that $x$ falls under the pattern in question.

(ii) You ask “What is ‘availability’?” and produce three possibilities. Roughly, I mean the sense you dub the strong sense (the quibbles aren’t worth the space).

(iii) You ask “what kind of sentences have to be available” and make the entirely correct point that we do have in the case of grammar a sentence available that represents that $S$ is grammatical, namely, the sentence ‘$S$ is grammatical’. How then do I count our knowledge of grammar as a good example of an implicit theory? The answer is that we cannot produce a sentence that elucidates the pattern. We know that being grammatical is not sui generis; we know there’s a structure to being grammatical but we cannot give it in words.

Now it is my turn to be baffled. I cannot see why you find all this so hard. There is a structure to being a wff in logic. Some students who know that being a wff is a structured property can give the structure in words (there’s more than one way to do this but that’s fine). Some cannot but they can reliably recognize wffs and they know case by case why a non-wff is a non-wff. We need some way of describing the knowledge of the second group. I say they have implicit knowledge. Maybe there are better terms; maybe there are better ways of making the distinction, but I cannot see why you resist the point that there is an interesting phenomenon here that calls for a name.
Maybe the answer lies in what you say about my talk of patterns but here I am equally baffled by your remarks. Anyhow, let me go through the cases that puzzle you in the section of your previous letter after Q15, “What are ‘patterns’?”: “Now it is surely true that, presented with some cases of ‘crook’ sentences, you know how to fix them. But, if ‘crook’ means ungrammatical, then we think it is simply false that, for each crook sentence, you know how to fix it. Every sentence in Hungarian is an ungrammatical sentence in English. But presented with a sentence in Hungarian and asked to ‘fix it,’ we would be stymied.” I don’t think that this last claim is true. Here’s one way to fix it. Put quote marks around it and add ‘is not a sentence of English’ to the result. That will do the trick. What is true is not that we don’t know how to fix it but rather that there is no salient candidate to be the way which deviates the least from what we started with.

You go on to say: “Perhaps you want to restrict your claim to ‘crook sentences’ that are made up of English words. But even here, we think your claim is false. It is easy enough to ‘fix’ your example, ‘She are happy.’ But how would you ‘fix’ something like ‘The the the in loud a a’?” (ibid.). Many answers: put it in quotes and append ‘is not well formed in English’; replace it by ‘The sound is loud’; etc.

You discuss (following Q15 in K&S[3]) my claim saying:

For any plane figure they [certain children] can say whether or not it is a circle and why—“It’s that bump that rules it out.” Here again, we’re puzzled. There are no doubt lots of kids who can do just what you say for figures like this one:

But what is the kid supposed to say when asked why the following figure is not a circle?
If we were asked this question, we would not have a clue about how to answer (unless: “Because it isn’t round!” counts as an answer).

But it is easy for the child to say why the figure is not a circle. What would puzzle them is a request to pick out the one reason—there are lots and lots of reasons. The bent bit at the bottom right, the bent bit at the top left, the flat section below the wave-like bit, etc., etc.

Why is it that we seem to so often to puzzle each other? Maybe the nub of it is in your question “But what could it mean to say that there is a pattern that unifies all the grammatical sentence of English?” (ibid.).

You go on to discuss this as if it were a question about language. For example, you discuss the idea that it means “the set of grammatical sentences must be recursive (or, a bit more modestly, that it must be possible to characterize them with a finite set of rules or principles)” (ibid.). For me it isn’t. If there’s a pattern, it will in principle be possible to capture it in words and there will be many ways to do this. But the existence of a pattern is a fact about the world—in particular about a certain set of sentences—and not about words. Square things are alike; we can capture the fact in words; but the similarity is not a fact about words. The difference between square things and grammatical sentences is the extent to which the similarity is disjunctive but the sense in which both are cases of things falling under a pattern is the same.

This means I find your puzzlement about my talk of driving itself puzzling. If there’s a pattern, why should we not be able to recognize it, and why shouldn’t the ability to recognize it do some driving—that is, some causing of sentences like ‘The sentence before me is grammatical’?

Or take the following passage from your letter (ibid.): “Finally, as we noted in our previous letter (K&S[2]) if there is one finite set of rules that generates a given infinite set, there are indefinitely many. As far as we can see, these are all the facts there are in this vicinity (as Dave Chalmers likes to say). So when you talk about a pattern that ‘unifies’ the grammatical sentences in English, which of these facts do you have in mind?” For me there is a glaring omission in the list of facts in this vicinity—you haven’t mentioned the pattern in nature. You’ve left out the main player, as I see things.

All the best,

Frank

K&S[4]: Letter from Kelby Mason and Steve Stich [4], July 11, 2005

Dear Frank,

Many thanks for your letter of 21 June. It’s been rough going. But we think we have finally begun to understand how you use the implicit/explicit
distinction, though there are still a number of issues on which we need lots of help. Like you, we continue to be surprised at how hard it is for philosophers who work in, near enough, the same philosophical tradition to understand each other. Sometimes we worry that the problem is idiosyncratic to us, and that we’re just being dense. If that’s the case then this exchange will be of very limited interest. But if, as we suspect, lots of others have seriously misunderstood your claims about folk psychology and implicit theories, then the time invested on this project will be well spent. In this letter, we’ll start with the bits we think we understand, then go on to the issues on which we still need help. In the last section, we’ll (finally) get back to folk psychology.

1 The Parts We Think We Understand

It was the following rather exasperated paragraph in your letter of 21 June (FJ[3]) that finally made your use of ‘implicit knowledge’ start to come into focus. (The numbering and emphasis is ours, of course.)

[A] Now it is my turn to be baffled. I cannot see why you find all this so hard. There is a structure to being a wff in logic. Some students who know that being a wff is a structured property can give the structure in words (there’s more than one way to do this but that’s fine). (iii) Some cannot but (i) they can reliably recognize wffs and (ii) they know case by case why a non-wff is a non-wff. We need some way of describing the knowledge of the second group. I say they have implicit knowledge. Maybe there are better terms; maybe there are better ways of making the distinction, but I cannot see why you resist the point that there is an interesting phenomenon here that calls for a name.

You are certainly right that the sentence we’ve emphasized characterizes an interesting phenomenon, that it calls for a name, and calling it ‘implicit knowledge’ is as good a label as any. But since we’ve misunderstood each other so often in the past, it’s important to go very slowly here and be very clear on what the label is a label for. As we read you, it is a label for an ability or capacity that some people have. And, as we’ve indicated with our numbering, the italicized sentence suggests that the ability has three components.

(i) The ability to recognize wffs—which, we assume, is the ability to tell, for any symbol sequence whether or not it is a wff. (Obviously, this is an idealization, since no one is very good at recognizing wffs that are, say, over ten million symbols long. And there are issues about how the idealization should be unpacked. But we don’t think that’s worth pursuing here, since there are similar issues about how to characterize just about any open-ended capacity.)
(ii) The ability to say, for each symbol sequence that is not a wff, why it is not a wff. (You talk about knowing rather than the ability to say, and these are surely different, though for the purposes at hand, we don’t think the difference will be important.)

(iii) Lack of the ability to “give the structure in words.”

If S has these three abilities (or, if you prefer, two abilities and an inability) you’d say: ‘S has implicit knowledge of wffs’, or perhaps you would prefer locutions like: ‘S implicitly knows what wffs are’ or ‘S implicitly knows what it is to be a wff’. We don’t think anything hangs on which of these locutions you’d use. But if you think there is some principled reason to prefer one of them, or some other similar locution, please let us know.

Though the account we’ve just given was couched in terms of the wff example, it is easy to see how to generalize to many other cases. The most obvious is the one we’ve discussed in previous letters, viz. grammaticality. If S (i) has the ability to recognize grammatical sentences in English, (ii) has the ability to say, for each nonsentence, why it is not a sentence, and (iii) lacks the ability to say what it is to be grammatical, then S has implicit knowledge of grammaticality in English (or implicitly knows what it is for a sentence to be grammatical in English). The generalization to another one of your examples, viz. circle, is straightforward as well. How to generalize to other important cases, including folk psychology, is less clear. But that’s a topic for Part 2.

Have we got it right so far? Or close to right? (We certainly hope the answer is yes. It would be a real bummer to have to start all over again!)

If we have got it right, then we can set to rest one concern that loomed large in previous letters. We have repeatedly protested that we did not understand your claim that an implicit theory “drives our classifications.” But as we now understand, for you, a crucial part of having implicit knowledge is having an ability to recognize. And, as you put it in FJ[3]: “If there’s a pattern, why should we not be able to recognize it, and why shouldn’t the ability to recognize it do some driving—that is, some causing of sentences like ‘The sentence before me is grammatical’?” (Emphasis is ours.)

Later on, we’ll have another go at the pattern-talk, which we take to be deeply problematic. But that’s not crucial here. What is crucial is that when you talk about implicit knowledge driving judgments or utterances, all you are claiming is that an ability to recognize grammatical sentences, for example, plays a causal role in producing the judgment that a specific sentence is grammatical. Of course it does. While this isn’t a very deep or particularly illuminating explanation of our grammaticality judgments,
you never said it was. But does it make sense to say that the ability to recognize grammatical sentences drives (= plays a central role in causing) grammaticality judgments? Of course it does. Issue resolved.

2 The Parts on Which We Still Need Help

In the previous section we suggested that the interesting phenomenon that you call ‘implicit knowledge of Xs’ can be understood as the conjunction of two abilities and an inability. We’ve got no problem with the first of these—the ability to tell, for any object, whether or not it is an X. (As we noted, there are idealization issues, but we’re not going to worry about those.) We do, however, have problems with both (ii) and (iii). We’ve tried to articulate the problems in previous letters, where we had a much less clear view of the role they play in your account of implicit knowledge; perhaps we can do a better job here.

2.1 Let’s start with (ii): knowing (or being able to say), for each non-X, why it isn’t an X. One thing that your most recent letter made clear is that for some reason this condition is important to you. In our last letter, we said we weren’t sure how important it was to get clear on this stuff about knowing (or saying) why a non-X isn’t an X (K&S[3]). But the issue has come up in the last two of your letters, and you spend quite a lot of time on it in FJ[3]. So clearly you do take it to be important. Unfortunately, we are still deeply puzzled about what this condition on implicit knowledge requires, and about why you think it is important to include it in your account of implicit knowledge.

To try to explain our puzzlement, let’s go back to the first time this condition (or something in the vicinity) made its appearance in our correspondence. Discussing the example of grammar, in FJ[2], you wrote: “[B] I know case by case what makes a sentence fall under, or fail to fall under, the pattern. I am not in the position of someone who says ‘I know “She are happy” is crook but search me how to fix it.’ I can say case by case for each crook sentence how to fix it (there will be many ways of course) and that case-by-case information is enough to construct in principle the sentential representation.”

Let’s assume, for the time being, that this earlier passage is spelling out (ii) as the ability to “fix” non-Xs. In K&S[3], we protested that while it is reasonably clear what would count as fixing in the case of near misses—almost-Xs like ‘the cat sitted on the mat’—there are plenty of ungrammatical sentences where we have no idea what fixing them would be. The examples we offered were a sentence in Hungarian and the word sequence ‘The the
the in loud a a', neither of which is grammatical in English. Asked to “fix” them, we insisted, we wouldn’t know what to say. In your most recent letter (FJ[3]) you respond to both cases. Regarding the Hungarian sentence, you write: “Here’s one way to fix it. Put quote marks around it and add ‘is not a sentence of English’ to the result. That will do the trick.” And for “The the the in loud a a” you suggest: “Many answers: put it in quotes and append ‘is not well formed in English’; replace it by ‘The sound is loud’; etc.”

Now you are certainly right that each of these strategies will produce a grammatical English sentence. What we find puzzling is the claim that these strategies will “fix” the ungrammatical sentence. You are, of course, free to use the term ‘fix’ however you see fit. But if you mean these examples seriously, then it seems that, as you use the term ‘fix’, there are just about no constraints on what counts as fixing. Is it enough to produce a grammatical sentence which either uses or mentions one or more of the words in the original ungrammatical sentence? That seems to be the only obvious feature that your three proposed fixes have in common. If that is enough to count as a “fix,” then the requirement that one must be able, on a case by case basis, to “fix” every ungrammatical sentence turns out to be absurdly weak, and it is hard to see why you want to impose it at all or what work it can do for you. To make the point in another way, consider the last bit of quote [B] above. If fixing is interpreted in the extremely weak way that your examples suggest, why would you think that the information obtained from fixes (in that weak sense) “is enough to construct in principle the sentential representation”? Frankly, this claim just makes our heads spin.

Now in quote [A], which may be a more careful statement of your view, you do not actually impose the able-to-fix-it requirement. So perhaps the able-to-fix-it requirement of [B] isn’t really a part of condition (ii). Rather, what you strictly require in [A] is (ii), knowledge (or the ability to say), for each non-X, why it isn’t an X. But here we’re back to our old problem. If we interpret this requirement in what we take to be the natural way, then it is absurdly strong. If someone gave us a Hungarian sentence and asked us to say why it is not a grammatical sentence in English, we would have no idea what to say (other than “because it isn’t English”). Similarly, if someone gave us the sentence ‘The the the in loud a a’ and asked us to say why it is not a grammatical sentence in English, we would not know what to say. So, interpreting your requirement in what we take to be the most obvious way, it would follow that Kelby and Steve do not have implicit knowledge of grammaticality in English. And this surely is not what you want to say.
There are, of course, lots of other ways in which requirement (ii) could be interpreted. We’re not at all sure that you have a clear interpretation in mind. But if you do have something clear in mind, it will have to avoid the twin pitfalls we’ve noted. It can’t be so weak (Give me a good sentence that uses or mentions one of the words in the bad one) that it excludes almost nothing, nor can it be so strong that it excludes folks like us from having implicit knowledge of English grammar.

Let us close this section with a pair of questions (well, OK, Q17 is actually a whole bunch of related questions):

Q16  How, exactly, do you propose to interpret requirement (ii)?

Q17  Why do you want to impose a requirement like this in your account of implicit knowledge? What work is it doing for you? Another way of putting this question is: Why don’t you just drop requirement (ii) and say that a person has implicit knowledge of X when he has the ability to reliably recognize Xs—as this idea is unpacked in (i)? (We’re ignoring (iii) for a moment; we’ll get to that next.) We suspect this is the crucial question, since what really puzzles us is why you think you need a requirement like (ii) in addition to (i).

2.2  Let’s turn, now, to requirement (iii). Here, in contrast with (ii), we think we understand why you want to impose some condition like this, since what this requirement is trying to ensure is that S does not have explicit knowledge. If S has the ability to recognize Xs and can also give a suitably detailed account of what Xs are, then you want to say S has explicit knowledge of Xs, not implicit knowledge.

But while the motivation is clear, the requirement, unfortunately, is not. We went on at some length about this in K&S[3]. What that discussion led up to was Q14, about the constraints on “belief capturing” open sentences. In your reply (FJ[3]) you never answered this question. The closest you come is to talk about a sentence that “gives the pattern.” But for reasons we’ll discuss in the section on patterns, we don’t think this even begins to count as an answer to the question.

2.2.1 A modest proposal  Up until now, in this correspondence, we’ve mostly been asking questions and trying to understand your view. But we now think we understand your view and its shortcomings well enough to propose an alternative. As we interpret your view, particularly the motivation for (iii), you want implicit knowledge and explicit knowledge to exclude one another. A person can have one or the other sort of
knowledge of a theory or pattern or property, but not both. But in order to draw the distinction, you need some clear and well-motivated answer to Q14. And, to be frank, we don’t believe that an answer will be forthcoming.

What we propose is a rather different way of thinking about the interesting (and not so interesting) phenomena in this vicinity. There are at least three of them.

(1) The first (basically (i)) is the ability to reliably recognize whether or not something is an X—where X can be grammatical sentence in English, wff, circle, and lots of other things.

(2) Second, someone who can recognize Xs can also know how he does it. In some cases (grammar, perhaps) this will involve knowing that his mind–brain uses a specific set of rules or principles to compute the answer to the question ‘Is this an X?’ One of the goals of much cognitive science is to produce knowledge of this sort, both about one’s own recognition abilities and about other people’s. And while there has been impressive progress in a few areas, there is still much work to do. There is very little of this sort of knowledge in the world at the moment. Perhaps none at all.

It is important to note that (1) and (2) are not exclusive. A person can have both (1) and (2). Moreover, even if someone has the sort of knowledge required for (2), she often will not use it to recognize Xs. If an English-speaking grammarian or psycholinguist ever does figure out how we distinguish sentences that are grammatical in English from sentences that aren’t, she will rarely if ever use this knowledge to recognize grammatical sentences in English. Rather, she will go on recognizing them the way she did before she acquired this new knowledge.

(3) Someone who can recognize Xs can sometimes know and state something interesting about what Xs are, or about the class of Xs, or about what all Xs share. This is a vague characterization and it covers a lot of ground. In the case of grammar, for example, it runs the gamut from being able to specify what Chomskians call a descriptively adequate grammar (see K&S[3]), to being able to state some or all of the rules that might be found in a traditional grammar book for English, to being able to state some even less explicit characterization of the class of English sentences (like those proposed in K&S [3]).

When you talk about “explicit knowledge” of X, it seems to be knowledge in this vicinity that you have in mind. But since saying something interesting about what Xs are is both vague and open ended, we don’t think that your notion of explicit knowledge is clear or interesting or
useful. Perhaps we’re being unfair here. Perhaps you do have some clear idea about what \( S \) needs to be able to specify about \( Xs \) if \( S \) is to count as having explicit knowledge about them. Perhaps, that is, you do have a clear answer to Q14, and perhaps that answer makes explicit knowledge an interesting and important notion. But we’re skeptical. The ball is in your court here.

At this point, we imagine, you might be thinking: “What’s the big mystery? It’s all to do with patterns. We know that there must be a pattern that unifies all the grammatical sentences of English. And to have explicit knowledge, a person has to be able to (as you say in FJ[3]) ‘produce a sentence that gives the pattern.’ To say why we find this unsatisfactory, we’ll have to once again take up the question we raised in K&S[3], namely. . . .

2.3 What are patterns? In addressing this question in your most recent letter (FJ[3]) you suggest what may be the nub of the problem, and we think you are spot on.

[C] Why is it that we seem to so often to puzzle each other? Maybe the nub of it is in your question “But what could it mean to say that there is a pattern that unifies all the grammatical sentence of English?”

You go on to discuss this as if it were a question about language. For example, you discuss the idea that it means “the set of grammatical sentences must be recursive (or, a bit more modestly, that it must be possible to characterize them with a finite set of rules or principles).” For me it isn’t. If there’s a pattern, it will in principle be possible to capture it in words and there will be many ways to do this. But the existence of a pattern is a fact about the world—in particular about a certain set of sentences—and not about words. Square things are alike; we can capture the fact in words; but the similarity is not a fact about words. The difference between square things and grammatical sentences is the extent to which the similarity is disjunctive but the sense in which both are cases of things falling under a pattern is the same. (FJ[3])

Here’s our diagnosis of the problem: We believe that most of this talk of patterns and similarity and structure and things being alike is (something like) a metaphor. If interpreted in a natural (and nonmetaphorical) way, it is simply false that all grammatical sentences are alike, or that they are all similar. Moreover, we do not think there is any good way of unpacking the metaphor. “Philosophy,” as Wittgenstein famously proclaimed, “is a battle against the bewitchment of our intelligence by means of language.” And we think you’ve been bewitched.

In an effort to break the spell, we’ll set a cluster of examples aimed at making it clear why we don’t think the notion of a pattern makes any clear sense when talking about grammatical sentences in English, or in many,
many other cases where you seem to think the pattern metaphor is unproblematic. Each one of our examples will be an artificial language whose sentences are sequences of standard Arabic numerals, of arbitrary length. So, for example, any of the following might be a sentence in one of our languages:

1221
00000500000009999999999999
17
1234567
90823758302757684736586878945352759

(Of course, nothing turns on calling these “artificial languages” or on calling the numeral sequences “sentences.” You can, if you prefer, just think of them as sets of numeral sequences.)

Language 1 is an infinite set of numeral sequences of arbitrary length; the members of the set are completely random. (We like to imagine having some device for randomly generating numeral sequences of arbitrary length, and just letting the device run for ever. But that’s window dressing.)

Language 2 is a finite language. It is formed from Language 1 by taking the first 1,000 numeral sequences. Those “sentences” and only those are grammatical in Language 2.

Language 3 is another infinite language. It includes ‘17’, ‘34’, ‘51’, and all the other numeral sequences that are (standard names of) multiples of 17.

Language 4 is the union of Language 2 and Language 3.

OK. Now let’s ask: As you use the notion of a pattern, in which of these languages is there a pattern to be found? We expect that 1 and 3 are the easy cases. The sentences in 1 were chosen at random, so we assume you want to say there is no pattern there. Indeed, if you think there is a pattern there, then your notion of a pattern is vacuous; it excludes nothing. We assume that you would say there is a pattern in Language 3, since the sentences in L3 all have something in common; they are all multiples of 17.

Now let’s turn to the hard cases. Is there a pattern in Language 2? Here, we think, different considerations pull in different directions. The sentences in L2 were chosen at random. And that surely suggests that there is no pattern there. On the other hand, it would be possible to learn to recognize sentences in L2 after a finite number of presentations, and in FJ[2] you say: “We know there must be a pattern that unifies the gram-
matical sentences in English, otherwise we could not acquire the ability to recognise the grammatical sentences in English after a finite number of presentations. Ditto for Russian etc.” Well, ditto for Language 2, too. So there “must be a pattern.” But now it looks like you are committed to the view that there is a pattern in any finite set of objects. So, given the things you say about patterns in quote [C] above, it looks like you are committed to saying that the objects in any finite set are “alike” or “similar” though perhaps the similarity they share is “disjunctive.” And if you do say that, then we simply have no idea what you mean by ‘alike’ and ‘similar’.

At this point, you might protest: “Of course there is no pattern in the first 1,000 elements of a set of numeral sequences chosen at random. It is the finiteness of Language 2 that’s responsible for the fact that it’s possible to learn it after a finite number of presentations, but the learnability test was never meant to apply to finite sets.”

But then what will you say about Language 4? Since it includes L2, if there is no pattern in L2 it’s hard to see how there could be a pattern in L4. But, like both L2 and L3, it is possible to learn to recognize sentences in L4 after a finite number of presentations, and L4 is infinite. So is there a pattern, or isn’t there? We have no idea how to answer this question. And the reason we don’t is that your notion of pattern is seriously underspecified. There is, we believe, no clear way to extend the pattern metaphor to interesting cases like language. (“See how high the seas of language run here.”)

2.4 How can your account of implicit knowledge be generalized to folk psychology? Before starting on this topic, we’d do well to summarize the state of play. If we’ve got you right, then when you say that S has implicit knowledge of X (where X can be grammar, wffs, circles, etc.) you are attributing a complex ability or capacity to S. The capacity has three features:

(i) an ability to recognize Xs;
(ii) an ability to say of non-Xs why they are non-Xs;
(iii) an inability to say what Xs are explicitly.

Though we have problems with (ii) and (iii), we take (i) to be relatively unproblematic. And since (i) will presumably be crucial in any application of the notion of implicit knowledge to folk psychology, that’s where we propose to focus our attention.

When we started this exchange of letters our goal was to better understand your conception of folk psychology. The way we got involved in the extended discussion of implicit theories was that, in your first letter, you
made it clear that you think folk psychology is an implicit theory. You also introduced the analogy with grammar. Here is the relevant passage:

Defenders of folk psychology often say that it is an implicit theory. I do. But what I mean by this is one, but only one, of the things sometimes meant when it is said that we have an implicit theory of grammar. There is an implicit theory that drives our classifications of sentences in languages we have mastered into the set of the acceptable and the set of the nonacceptable sentences. This is the theory we make explicit by interrogating our intuitive classifications and which, when extracted and recorded in words, makes its way into grammar books as an explicit theory of grammar. That’s how grammar books get written. (FJ[1])

Now that we’ve made some progress in understanding what you mean by ‘implicit theory’ (or at least we hope we have) we’ve started to think about what it might mean to say that folk psychology is an implicit theory, focusing on (i), the recognition component of implicit theories. And we need some help. Presumably, on your view, if a person has implicit knowledge of folk psychology, she has a complex ability which includes an ability to recognize something. What we are not clear on is what the something is. The problem is not that there are no candidates. Quite the opposite, there are lots of candidates, some more problematic than others.

In “What Is Folk Psychology?” Stich and Ravenscroft (1996, 124) assembled “a partial list of the ‘folk psychological’ capacities or abilities that need explaining.” These included the following (the numbering here is not the same as the numbering in the article):

1. The ability to attribute beliefs, desires, emotions, and other mental states to oneself and to others.
2. The ability to predict how people will behave.
3. The ability to construct explanations of people’s behavior couched in mentalistic terms.
4. The ability to judge the correctness or incorrectness of psychological principles or generalizations couched in mentalistic terms. (What we had in mind here were Lewis-style “platitudes” like “When a normal person is looking at a traffic light which changes from red to green, she usually comes to believe that it has changed from red to green.”)

More recently, Shaun Nichols and Steve published a book, Mindreading (Nichols and Stich 2003), in which they pointed to a number of other abilities that the folk have which need explaining. Here are some drawn from the list on page 77:

5. The ability to predict the inferences that other people will draw.
6. The ability to attribute a limited number of perceptual states to other people. (It turns out that there are surprising shortcomings here.)
7. The ability to predict some of the decisions that other people will make. (Here again there are surprising shortcomings.)

Now, if one is not too fussy, one might describe most or all of these as recognition abilities. (1) I can recognize when $S$ believes that $p$, and when $S$ does not. (2) I can recognize when $S$ will call the police and when $S$ won’t. (4) I can recognize psychological generalizations (“platitudes”) that are true (or plausible, or sound right) and generalizations that are not. And so on.

As noted in the Stich and Ravenscroft article, the grammar analogy works best for (4)—judging the correctness of generalizations. Just as our intuition tells us that some sentences are OK and others are crook (as you so delightfully put it), so too our intuition tells us that some putative platitudes are OK and others are crook. So perhaps when you say that people have implicit knowledge of folk psychology, what you mean is just that they have the ability to recognize good and crook putative platitudes, and all the rest of these abilities are irrelevant. On the other hand, perhaps you would want to include some or all of these other abilities among the recognition abilities that a person must have if she has implicit knowledge of folk psychology. There are, we think, some very sticky issues to be faced if you go the latter route. But there is no point in exploring those until we know which way you go here. And besides, this letter is already way too long.

All the best,
Kelby and Steve

FJ[4]: Letter from Frank Jackson [4], August 4, 2005

Dear Steve and Kelby,

The two key things I need to say to assist with our mutual comprehension problem is, first, I do think any finite set of items automatically exemplifies a pattern, and, second, without clause (ii) we would not necessarily have a case of (implicit) knowledge of what it takes to be a wff, say. Let me say something about both points in turn.

1 Patterns
My attitude to patterns is like Quine’s to objects. Very inclusive. The city of London plus the coin on my desk is an object. Not one anyone would
have much interest in—except to make a point in a philosophy seminar, and not one there is a word for in English though we could introduce one, but an object all the same. Ditto for patterns. Thus any finite sequence of numbers exemplifies at least one pattern. Actually if we are dealing with a sequence of word tokens for numbers like the sequences in your letter, there will be a huge number of patterns: being written on a certain page, being produced by such and such a person, being of so-and-so a temperature at such-and-such a time, etc. Of course the set of English sentences is not finite and the pattern that unites the grammatical ones is reasonably unified and of interest—indeed some importance in everyday life. Ditto for wffs except the interest is more limited.

This means that saying that there exists a pattern in and of itself is typically a very weak claim. What isn’t a weak claim or need not be is saying that so-and-so a pattern is what some bit of English picks out, or saying that some pattern or other is something implicitly or explicitly known.

2 Clause (ii)
You can have the ability to classify formulas of logic into wffs and non-wffs in the absence of an ability to give the, or a, general formula that covers all cases without having an implicit theory of wff-ness. This is why clause (ii) is needed. Think of a student a bit like *Clever Hans*. The student correctly classifies formulas into wffs and non-wffs but is doing it—unwittingly let’s suppose—by noting something about the expression on the tutor’s face when she, the tutor, looks at any given formula. This student does not have implicit knowledge of what it takes to be a wff.

The key point about clause (ii) is that it means that we could construct a theory of grammar or wff-ness in the explicit sense from the knowledge of the person who only has an implicit theory of grammar or wff-ness. Indeed that is how explicit grammars were constructed and how many of us came to have an explicit theory of wff-ness. The case-by-case knowledge is enough to allow us to find the words that cover all the cases.

Hope this helps,
Frank

*K&S[5]: Letter from Kelby Mason and Steve Stich [5], September 11, 2005*

Hi Frank!

One issue that needs to be addressed is where to end the discussion. In the last section of our letter of K&S[4] we raised some questions about how
your account of implicit theories applies to the case of folk psychology. However, you did not address those questions in FJ[4]. We’re inclined to think that the article might be of greater interest if it included some discussion of this issue. So, if you are not yet tired of the exchange, we’d like to encourage you to say something about it.

All the best,
Steve and Kelby

---

**FJ[5]: Letter from Frank Jackson [5], November 14, 2005**

Dear Steve and Kelby,

I fear we are still at cross-purposes on many issues. I’ve tried to be as clear as I can in my response.

1 **What Do I Mean by Saying That S has an ‘Implicit Theory’?**

I mean S has a theory which is not explicit. I’m not ‘attributing a complex ability or capacity’ to S, though S may have—typically will have—various complex abilities.

What’s a theory that is not explicit for S? It is one (i) S holds but (ii) S cannot give the content in words. If S is asked on an exam to state the content of theory T, when S only knows T implicitly, S fails that question.

To have a theory is to have a certain view about how things are, and so any philosopher who thinks that one can have views about how things are which cannot be put into words should believe in implicit theories.

The material on recognition in my discussions of implicit knowledge and theories is not part of what it is to be implicit. The reason for giving the examples where we can recognize that something is a T (to be a T is to satisfy T) is because they are a good source of cases where subjects have an implicit theory. However, recognitional ability is not part of what it is to have an implicit theory.

Here is an example where a theory is implicit but there is no recognitional ability. Fred is ‘motion blind’ in the sense that he cannot see things as moving. However, whenever given information like:

Mary is in Melbourne at $t_1$; Mary is in Sydney at $t_2$; Melbourne and Sydney are in different places

Fred infers that Mary has moved.

However, he cannot produce anything like ‘x moves iff x is in different places at different times’ when asked for the conditions under which
something counts as moving. He has, that is, two deficits by comparison with most of us. He cannot see things as moving, and he cannot give the rubric.

He has (say I but I think this is also what Lewis and many would say) an implicit grasp of what motion is; he has an implicit theory of motion.

2 So How Does My Account of Implicit Knowledge Generalize to Folk Psychology—Your Question?

It depends on what you mean by ‘folk psychology’.

(i) If you mean anything the folk mostly hold about psychological states, there are many things they believe which are not implicit. Take:

Pain is unpleasant.

This is something the folk believe and they can put it in words easily enough.

(ii) There is, however, a view about the mind associated most especially with Lewis and analytical functionalism more generally (I know there are ways of carving up the territory that excludes Lewis from the class of analytical functionalists but here they belong together) that can be sketched as follows.

There is a theory $P$ such that

(1) it is held by the folk;
(2) it is largely implicit (for the folk);
(3) a subject is in mental state $M$ iff the subject satisfies the open sentence for $M$ that comes from $P$ in the usual way.

Those who hold this position often call $P$ ‘folk psychology’. $P$ is not everything the folk believe about the mind. It is the theory that satisfies (1) to (3) above, if such there be.

It is on the second sense of ‘folk psychology’—the sense on which folk psychology is that which satisfies (1) to (3) above—that the notion of an implicit theory is important. Indeed it is crucial for the plausibility of the view about the mind just outlined. This is because it is implausible that there is an explicit theory that satisfies (1) to (3) above. In other words it is implausible that:

There is a theory $P$ such that

(1) it is held by the folk;
(2*) it is explicit (for the folk);
(3) a subject is in mental state $M$ iff the subject satisfies the open sentence for $M$ that comes from $P$ in the usual way.

However, nothing I say about what an implicit theory is supports the above view of mind. One might well hold that there are implicit theories while rejecting the Lewisian view of the mind sketched above.

I am wondering if the misunderstanding between us arose because you expected more from the notion of an implicit theory? Perhaps you thought the notion was in itself an argument for the above view of the mind?

Best,

Frank

K&S[6]: Letter from Kelby Mason and Steve Stich [6], April 15, 2006

Results! Why man, I have gotten a lot of results. I know several thousand things that won’t work.

—Thomas Alva Edison, who did most of his work just a few miles from Rutgers

Dear Frank,

Back in July, when we sent off K&S[4], we thought we were making real progress in understanding your view. But your two most recent letters, FJ[4] and FJ[5], make it clear that we were unduly optimistic. We’re still almost totally unable to understand your view, and we suspect that you are equally puzzled about why we don’t understand you. Since this must be our last letter, what we’ll do here is offer our take on the state of play in this dialogue. For the most part, it will be a catalog of things we do not understand. Still, if Edison was right, we’ve gotten a lot of results, even if the lightbulb has only flickered now and then.

Let’s start with a brief reminder of what we thought we understood in K&S[4]. At the core of our mutual miscomprehension is your notion of implicit knowledge, and in K&S[4] we proposed an analysis of that notion. Implicit knowledge of $X$s (where $X$ can be wffs, grammatical sentences in English, circles, or any other “pattern”) consists of two abilities and an inability:

(i) the ability to recognize $X$s
(ii) the ability to say for each non-$X$ why it is not an $X$
(iii) lack of the ability to “produce an open sentence . . . that gives the pattern” FJ[2]

We went on to explain that we had concerns about (ii) and (iii), but that (i), at least, was more than clear enough to work with.
The most discouraging part of FJ[5], for us, is your insistence that we were wrong about (i). “[R]ecognitional ability,” you write, “is not part of what it is to have an implicit theory.” The context makes it clear that recognitional ability is not necessary for implicit knowledge, on your account; it is less clear whether you would say that conjoined with (ii) and (iii) it is not sufficient.

We suspect—and hope—that here we’re just talking past one another again. As we view the example you give, it sounds reasonable to say that motion-blind Fred does have a recognitional ability, albeit a highly inferential one. Given the relevant facts about Mary, he can recognize the case as one of motion. But it seems that you consider this as not recognitional, because Fred can’t perceive the motion. So perhaps you are interpreting “recognitional ability” as a kind of perceptual ability. If so, then perhaps you’d want to say that Fred has a classificatory ability but not a recognitional one. So let’s try rephrasing (i) thus:

(i’) the ability to recognize or classify Xs (as Xs)

We have our fingers crossed that you’d be willing to grant this kind of ability as a necessary part of implicit knowledge. If not, then we’re back to square one. If (i’) (or something in the vicinity) is not even part of what it is to have an implicit theory, then we have, near enough, no understanding at all of what you think implicit knowledge is.

In the same section of FJ[5] in which you say that recognitional ability is not part of implicit knowledge, you make one more attempt to say what you mean when you say that S has an implicit theory. “I mean S has a theory which is not explicit....” And “[t]o have a theory is to have a certain view about how things are, and so any philosopher who thinks that one can have views about how things are which cannot be put into words should believe in implicit theories.” One of the disadvantages of having this discussion via email rather than across a table is that it is sometimes hard to discern the “tone of voice” with which comments are made. The first time we read this, we imagined it written in exasperation. Surely you couldn’t think this would help us to understand your view, so you were probably just throwing up your hands and giving up on us. We still, mostly, think this is the way to read the remarks we’ve quoted. But from time to time we entertain the idea that you did think this might help because in your corner of the philosophical world “having a certain view about how things are” is an expression which is clear enough, and precise enough, to use in careful philosophical discussion. If so, then the divide between your corner of the philosophical world and ours is even greater.
than we had imagined. Of course we think that “one can have a view about how things are.” But in our philosophical dialect, the expression is far too vague to be of any help in understanding what you mean by ‘implicit theory’.

Let’s turn, now, from (i) to (ii). In K&S[4] we went on at some length about why we found your condition (ii) puzzling, and in FJ[5] there were two short paragraphs in which you expanded on your view. Unfortunately, we did not find them very helpful. Indeed, as far as we can see, they did not really address our concerns at all. Here’s how we see the state of play on (ii). You have articulated the requirement in several different ways. In FJ[2], where the example in question is implicit knowledge of grammatical sentences in English, you say: “I can say case by case for each crook sentence how to fix it” (emphasis ours). In FJ[3], where the example at hand is implicit knowledge of wffs, you say that those who have implicit knowledge “know case by case why a non-wff is a non-wff.” In both K&S[3] and K&S[4] we asked for clarification on how to interpret these remarks.

The problem, as we see it, is that on the natural interpretation it is simply false that the sorts of people to whom you would clearly attribute implicit knowledge of Xs know why non-Xs are non-Xs. Let’s focus on grammar, which has been center stage in much of our correspondence. Both of us are native English speakers, and it is clear that you would attribute to us an implicit knowledge of grammaticality in English (or an implicit theory of English grammar). Both of us recognize that the sequence of words ‘would brick that clear is and bit us to’ is not grammatical in English. But if asked why it is not grammatical in English, we would have no idea what to say. Nor would we know what to do if you asked us to “fix it.” In FJ[3] you say this is not true; you insist that we do know how to fix sentences like this. “Here’s one way to fix it. Put quote marks around it and add ‘is not a sentence of English’ to the result.” But as we note in K&S[4], this makes (ii) an extremely weak requirement. You are of course, free to interpret the requirement in any way you wish. But in several places, including most recently FJ[4], you have maintained that the sort of “case by case” knowledge required by (ii) is going to do important work for us. Your most explicit statement on this is the following in FJ[2]: “I can say case by case for each crook sentence how to fix it (there will be many ways of course) and that case-by-case information is enough to construct in principle the sentential representation” (emphasis added). But on the weak interpretation of (ii) suggested by your “put quote marks around it . . .” proposal, this strikes us as patently absurd. How could the information gleaned from that sort of “fixing” be of any use at all?
So as we see it, here’s the bottom line on (ii). Interpreted in what we take to be the natural way, it is just false that speakers typically know how to fix nonsentences or to say why they are not grammatical. So on this interpretation, speakers typically fail to satisfy (ii) and thus they do not have implicit knowledge of the grammar of their language. And that’s bad news for you. Interpreted in the weak “put quote marks around it” way, the information provided by speakers’ case by case “fixes” is utterly useless. And once again, that’s bad news for you. Clearly, you’ve got some work to do here. You need to say more explicitly what (ii) requires, and to explain why you think the case-by-case information that can be obtained from people who meet that requirement is of any use.

Let’s now move on to (iii). In K&S[2] we asked what we thought was a straightforward question:

Q14 What are the constraints on the “belief capturing” open sentence that must be available to a person if his theory is to count as explicit rather than implicit?

And in S&K[4] we protest that “you never answered this question. The closest you come is to talk about a sentence that ‘gives the pattern.’” We went on to say why we thought this was too vague to be of help:

Someone who can recognize Xs can sometimes know and state something interesting about what Xs are, or about the class of Xs, or about what all Xs share. This is a vague characterization and it covers a lot of ground. In the case of grammar, for example, it runs the gamut from being able to specify what Chomskians call a descriptively adequate grammar (see S&K[3]), to being able to state some or all of the rules that might be found in a traditional grammar book for English, to being able to state some even less explicit characterization of the class of English sentences (like those proposed in S&K[3]).

When you talk about “explicit knowledge” of X, it seems to be knowledge in this vicinity that you have in mind. But since saying something interesting about what Xs are is both vague and open ended, we don’t think that your notion of explicit knowledge is clear or interesting or useful.

In FJ[5] you take up the matter again: “What’s a theory that is not explicit for S? It is one (i) S holds but (ii) S cannot give the content in words. If S is asked on an exam to state the content of theory T, when S only knows T implicitly, S fails that question” (emphasis added). As we said earlier, in discussing the first part of this quote, we’re not sure what “tone of voice” to attribute to you here. Perhaps you are so exasperated by our questions that you can no longer take them seriously. Perhaps you mean the emphasized passage as a joke. If it is intended as a joke, let us respond in kind.
Standards on exams differ dramatically from place to place and from time to time, and in the United States, “grade inflation” has become rampant during the last few decades. Thus at many U.S. schools nothing would count as implicit knowledge, by your standards, because no one fails! It’s not a great joke, to put it mildly. But it does have a point. Your failing-the-exam account of implicitness is far too vague to be of help to us in trying to understand your view.

Finally, let us say something about the vexed topic of patterns. It was here that FJ[4] held its biggest surprise. Frankly, we never entertained the possibility that you would say that “any finite set of items automatically exemplifies a pattern.” And we are still struggling to come to grips with the implications of this revelation. Why were we so surprised? Well, in talking about patterns, both in this correspondence and elsewhere, you repeatedly use words like “similarity” and “alike” and “structure,” and you use squares and circles as standard examples of things that exhibit a pattern. Here is an entirely representative example from FJ[3]: “[T]he existence of a pattern is a fact about the world. . . . Square things are alike; we can capture the fact in words; but the similarity is not a fact about words. The difference between square things and grammatical sentences is the extent to which the similarity is disjunctive but the sense in which both are cases of things falling under a pattern is the same.”

Once you have made it explicit that, on your conception of pattern, any finite set of items exemplifies a pattern, all of this is, at best, seriously misleading. To emphasize the point, consider an example. Let \( R \) be a finite set of things chosen completely at random. For concreteness, imagine that they are chosen as follows. We launch a military drone on a series of flights around the world. Every second it snaps a photo, and analysts identify the most salient object in the center of the photo. Then a truly random physical process—one that depends on cosmic ray impacts, perhaps—is used to decide, for each salient object, whether or not it is in \( R \). When \( R \) contains exactly 982,475,893 members, the flights are terminated. On your account, the members of \( R \) exemplify a pattern. So readers of the above passage are encouraged to think that the members of \( R \), like square things, “are alike”—that there is some disjunctive “similarity” between them, and that the existence of the \( R \) “pattern,” and of the “pattern” associated with every other finite random set, is a “fact about the world.” Is it any wonder that we were misled? Can there be any doubt that most other readers would have been misled as well?

You acknowledge in FJ[4], that your “attitude to patterns is . . . [v]ery inclusive.” And we have been lamenting the fact that you were not more
explicit about this earlier in our correspondence and in your other writings. But now, unless we have misunderstood you yet again, your usage is clearer. For you, the members of every set fall under or exemplify a pattern, with the exception of infinite random sets. We are still not entirely clear about your conception of patterns, however. In FJ[4] you say that some finite sets, like our Language 2 (in K&S[4]), exemplify “a huge number of patterns.” If that’s not a slip, then some further clarification is needed.11

What puzzles us now is why you think this very inclusive notion of pattern is going to be of any interest in the philosophy of psychology or the philosophy of mind. Consider, for example, the sort of classificatory behavior that looms large in your account of conceptual analysis.12 Is it of any interest to know that an agent’s classificatory behavior—or that all the things she classifies as Xs—exhibit a pattern? As far as we can see, the answer is no, since we know in advance that this has to be the case. For suppose that someone actively tried to classify things as Xs in a way that Xs exhibited no pattern at all. She is bound to fail, since sooner or later she will die and the set of things she’s classified as Xs, whatever they are, will exhibit a pattern. Indeed, if your comment quoted at the end of the previous paragraph was not a slip, the things she’s picked out will exhibit a huge number of patterns. How these facts could be of interest to philosophers of mind or philosophers of psychology, or anyone else, remains a mystery.

We started this letter with Edison, so let’s finish with some words of wisdom from another great American inventor, the well-known epistemologist Donald Rumsfeld: “[T]here are known knowns; there are things we know we know. We also know there are known unknowns; that is to say we know there are some things we do not know. But there are also unknown unknowns—the ones we don’t know we don’t know.” We set out on this correspondence in the hope of better understanding your view, thereby expanding our own stock of known knowns. The notions of implicit theory and folk theory have been important in many philosophical debates over the last fifty years, and in much of your own work especially, so wouldn’t it useful if we could get clear about how you understand these notions?

So we thought, and so we still think. Alas, we’re a long way from there, even after our many exchanges. But we have managed one thing, and that’s to reduce our stock of unknown unknowns. At least now we know there’s lots about your view that we don’t know, and don’t understand. Which is some kind of progress, at least; even Rumsfeld might approve.

All the best,
Kelby and Steve
FJ[6]: Letter from Frank Jackson [6], November 6, 2006

Dear Steve and Kelby,

I too am sorry we seemed so often to be at cross-purposes but I doubt if giving a detailed set of responses to your summing up of 15 April 2006 (K&S[6]) would assist. Indeed, induction suggests the opposite. However, maybe two general remarks from me will be of value—perhaps for the three of us, or perhaps and hopefully for readers of the correspondence.

1. I want to say something fairly limited about the notion of an implicit theory; you were always reading me as saying something more far reaching. I hold that people can have a theory that things are thus and so without being able to put it into words. If I am right, we need a name for this kind of case and I think ‘implicit theory’ is not a bad name. I guess I still don’t know where you stand on the possibility.

I gave a number of examples where it seems to me that people hold a theory that things are thus and so without being able to put the theory into words. Each example had one or another special feature, as is the way with examples, and then—somehow—the discussion became one about those special features.

Take the well-worn example of students who can reliably recognize wffs in logic but cannot give a definition of a wff. When they say that some formula is a wff, surely there is a sense in which they are making a claim about how things are and so have theory about how things are which they are expressing when they that some formula is a wff. At the same time, they cannot produce something of the form: \( x \) is a wff if \( x \) is. . . . Well that isn’t quite right. They can no doubt produce sentences like: ‘\( x \) is a wff iff \( x \) is a wff’, and ‘\( x \) is s wff iff \( x \) is said to be a wff by an expert logic instructor’. What they cannot do is give an illuminating account of what makes a wff a wff. We need a way of describing this kind of situation. I say that they have an implicit theory of wff-hood.

One thing that baffled me about our correspondence is that at the end of it I still did know where you stood on this question. I knew you had worries about the notion of recognition and of the use of ‘theory’ but not how you yourselves would describe the kind of situation just sketched.

2. I want to distinguish a theory from the statement of a theory. Of course one says things like ‘\( E = mc^2 \)’ is part of STR, which might suggest that the statement itself is part of the theory. But, in my view, it is how things have to be for the statement to be true that is the theory, or the relevant part of the theory. I felt at a number of points in our exchanges that you were using ‘theory’ more for one or another statement of a theory,
than for a theory in the sense I had in mind. Of course one is free to use ‘theory’ for a statement-like animal, but then you would need another term, say, ‘theory*’, to discuss what I was discussing.

If a theory is not a statement, what kind of animal is it? I think it is a way things might be, and a true theory is a way things might be that obtains. ‘E = mc²’ contends that a certain way things might be obtains at the actual world. (What about inconsistent theories and mathematical theories—good questions for another time.) It is, that is, a view about what our world is like. This is the same as saying that it is a view about the patterns exemplified in the actual world.

You worry that I have a very inclusive notion of a pattern. But I don’t have a very inclusive notion of a pattern of interest, or a pattern affirmed by one or another theory, or . . . , so I was at a loss to understand what worried you so much about my inclusive notion of a pattern per se.

Best, as always,

Frank

Notes

5. Chomsky also required that these rules should entail many facts about the grammatical properties and relations of sentences in the language. But for present purposes we can ignore these complications.
6. It is the account developed in Stich and Ravenscroft 1996, sections 2 and 3.
7. Suppose there is someone who fails to meet the able-to-fix-it requirement. All we need do is teach him the put-it-in-quotes-and-append–“is not grammatical in English”-trick, and voilà, he can fix any ungrammatical sentence.
8. If you doubt this, let L2 contain the first 100 sentences in L1, rather than the first 1,000.
9. Note that although our example of L4 is both artificial and quite simple, it does have some important features in common with natural languages. Like L4, natural languages can be described by a bunch of recursive rules plus a finite bunch of apparently random exceptions and special cases and lexical rules. Pattern? No pattern? We have no idea, because we don’t know what is being asked.
10. Or, better, the recognition component of what you mean.

11. Do all finite sets exemplify a huge number of patterns? How about nonrandom infinite sets? Do they, too, exemplify a huge number of patterns? And just how huge is that huge number? Do some (or all) finite sets and nonrandom infinite sets exhibit an infinite number of patterns?

1 Introduction

When Jack Smart, David Lewis, and David Armstrong advanced their physicalist views of mind (see Smart 1959; Armstrong 1968; Lewis 1966 [in 1983]), they offered as part and parcel of their argument sentences of the following form:

\[ x \text{ is in } M \text{ iff } x \text{ is so and so} \]

where (i) ‘\(M\)’ comes from some standard list of mental state terms, so ‘\(M\)’ might be ‘pain’ or ‘desires tea’, (ii) ‘so and so’ is such that it is transparent in some sense that being so and so could obtain solely in virtue of the physical way the world is (examples are where ‘so and so’ is a functional or a topic-neutral specification noncontroversially fulfillable without remainder by the physical); and (iii) the biconditional is not an interesting discovery about what our world is like but instead is a priori true.

In doing this they were very much working within a tradition. To give one example among the dozens that might be given, the many who have defended compatibilist accounts of free action have by and large taken it for granted that part of the defense of compatibilism involved defending claims of the following form:

\[ x \text{ is a free action iff } x \text{ is thus and so} \]

where (i) ‘thus and so’ is such that it is transparent in some sense that being thus and so could obtain solely in virtue of the deterministic way a world is, and (ii) the biconditional is not an interesting discovery about what our world is like but instead is something claimed to be a priori true.

What is the rationale for this tradition? Why suppose that anything like the above needs to be done? This paper offers one answer to this question.
I have of course tried to answer this question in the past, in most detail in *From Metaphysics to Ethics* (Jackson 1998a). What I say here has some connections with what I say there but differs substantially. This is not because I wish to go back on what I said there but because I have become aware of other paths to the same general position.\(^1\)

There are five parts to the answer I will give: one part is a thesis about properties and similarity; one part is a thesis about similarity and logical space; one part is a thesis about certain claims in fundamental metaphysics; one part is a thesis about similarities between aggregations; and one part is a certain optimism about our ability to capture similarities and patterns in words. With that many balls in the air, there is ample scope for dissent somewhere along the way, but I hope that even the strong dissenters to one or another stage in the argument will find other stages attractive and will, in any case, find the ride interesting and illuminating. My aim is to make the argument path as clear as I can, not to overwhelm with the detail of the various considerations along the way.

### 2 Properties and Similarity

One thing we are doing when we do metaphysics is asking what our world is *like*, and part of the answer is an enumeration of the ways in which things are alike and different, one from another. Ways in which two people may be alike include: having minds, being rational, believing that snow is white, being made largely of water, desiring world peace, living in Paris, being taller than more than four people, and so on. We can think of the property of being rational, say, as the relevant similarity between all subjects that are rational. (I’m presuming a nonexpressivist position on rationality.) In saying it this way, I am making an ontic commitment to the property of being rational: it *is* the relevant similarity. Resemblance nominalists will argue that I could have said it in a way that avoided the ontic commitment, but nothing in what follows requires us to enter that debate. Behaviorism will then be one theory about what that similarity is, functionalism another theory, and the view that rationality is sui generis yet another. *Universals* in David Armstrong’s sense are very special ways of being similar that carve nature at its most fundamental joints; but similarities of any degree count as properties, although some will be boring ones of little interest. Being rational and mental properties in general are examples of interesting, valuable properties cum similarities that are obviously not universals; they do not carve nature at its most fundamental joints.
I have been asked in discussion: “Why think that there must be something similar about all the rational beings, all those who desire world peace and so on?” One reply is that we could not have learned how to use ‘rational’, ‘desires world peace’, and so on, unless there are similarities in the sense of projectible patterns. But a deeper reply is that our very use of these terms and these categories in our attempts to make sense of the world we live in, and to pass on information about what it is like using these terms and categories, reflects our conviction that there exist similarities corresponding to them—and the similarities are not exhausted by the fact that the terms apply and the items fall under the categories: the terms apply to the items and the items fall under the categories because of how the items are.

From this perspective, physicalism is a view about how those in, for example, pain are alike. They are alike in physical ways, and not in the kind of way envisioned by dualist theories of mind. The key difference between physicalists and dualists over consciousness is then over whether there is something alike between those who are conscious that outruns anything to be found in the picture the physical sciences give of our world. From this perspective, compatibilism about free will is a view about how free acts are alike that insists that the alikeness is one that can be found between items in worlds where determinism is true (which is of course consistent with holding that the alikeness obtains also in worlds where determinism is false, ours for probable example).

3 Similarities and Logical Space

Those who are in pain in our world are alike in being in pain. They are in addition alike in indefinitely many other ways. They are alike in being made of carbon, in being larger than a pin, in being $L$-located, where $L$ is the scattered location of all subjects that are in pain, and so on. Is there some way of winnowing out the “wrong” similarities? The possible worlds framework offers a way into this issue.

Take the familiar question of the relation between being an equiangular triangle and being an equilateral triangle: one property or two? In one sense there would seem to be two properties. Surely, in some sense one can believe that something is an equilateral triangle without believing that it is an equiangular triangle, and conversely. Isn’t that the situation of those who fail to realize that, necessarily, $x$ is an equiangular triangle if and only if it is an equilateral triangle? On the other hand, being an equiangular triangle is being a certain shape, and that shape is the very same
shape as being an equilateral triangle. A way out of this puzzle is to distinguish property concepts from properties (in re). Being an equilateral and being an equiangular triangle are the very same property, but we can represent it in language or thought as that which lies in the intersection of the triangles and the plane figures with all angles equal, or as that which lies in the intersection of the triangles and the plane figures with all sides equal. We can take the same approach for other pairs like: being half-full and being half-empty, being taller than the person next to you and having the person next to you being shorter than you, and being a plane figure each point of which is equidistant from a single point and being a plane figure taking the maximum value of area to perimeter. It seems implausibly profligate to posit distinct properties in cases like these. What we should say in each case is something like that there is only one property or similarity in nature while granting that there are two property concepts.

If this moderate ontology of properties or similarities is correct, then we can do the winnowing out by following the lead of David Lewis and Robert Stalnaker (see chapter 1 of Lewis 1986a, and Stalnaker 1984). We look to the similarities across logical space. Being in pain is the single similarity that unites all subjects in pain in logical space; likewise for consciousness, free action, and all the rest. When we ask after the nature of free action, consciousness, inflation, pain, desire for tea, and so on and so forth, we are asking after that which unifies, the similarity that holds of, all free actions, conscious states, cases of inflation, pains, and desires for tea in logical space.

4 Aggregation and Fundamental Metaphysics

What is it to hold that a house is nothing more than certain bricks, bits of mortar, glass panes, lengths of timber, sections of paint, and so on? A natural answer appeals to mereology. It is to hold that the house is the mereological sum of those parts. It is simply the aggregation of those parts arranged in the way that they are. The ‘is’ here is the ‘is’ of constitution. As nearly all agree, the house differs in its persistence properties from the aggregation that makes it up. Some hold that recourse to a metaphysics of temporal parts allows us to construe the ‘is’ of constitution in terms of identity, but we can set that contentious issue aside; nothing in our brief turns on it.

It is natural to think of many claims about the fundamental nature of mind, free will, consciousness, our world or whatever, as claims about what can be got from the aggregation of such and such ingredients alone. An
obvious example is Humeanism about laws: it holds that suitable aggrega-
tions of matters of particular facts deliver in and of themselves laws. But
perhaps the most famous example is the claim that physicalism about
consciousness is the claim that God could make consciousness by putting
together right ingredients that come solely from those that feature in the
physical sciences—that there would be nothing more for God to do.
Setting the provenance aside, the idea is that physicalism is the claim that
creatures with minds are mere aggregations of the physical, where ‘mere’
signifies that are they nothing over and above such aggregations. Likewise,
compatibilism about free action can be thought of as holding that suitable
aggregations in deterministic worlds compose subjects that sometimes act
freely. Physicalism about our world—that is, physicalism about all of it and
not just the minds in it—is the view that our world is a huge aggregation
of the physical. Here we need, of course, to include the laws of nature as
part of what is aggregated (unless they supervene in Humean fashion), how
bits relate to one another, and force fields of various kinds. What is aggre-
gated is not merely ‘material bits’. In fact, the idea really is that our world
is a vast exercise in conjunction: electron $e_1$ at place $p_1$, and electron $e_2$
at place $p_2$, and $\ldots$ and force field $f_1$ distributed through region $r$ with curva-
ture $c_1$ and $\ldots$. The claim about fundamental nature boils down to a claim
about the nature of the conjuncts: physicalism about our world is the thesis
that each and every conjunct comes from the physical sciences. Mutatis
mutandis for other examples.

Our concern here is with theses in fundamental metaphysics (or ‘serious
metaphysics’ in the terms of Jackson 1998a, chapter 1) that can be thought
of as aggregation theses: theses about what can be got from what by (mere)
aggregation. How does this fit with the idea that properties are similarities
across logical space, and the idea that one topic in fundamental metaphys-
ics is (surely) the nature of a property like consciousness or acting freely?
Here we need to draw a distinction between metaphysics across worlds and
metaphysics at worlds.

5 Metaphysics at Worlds and Metaphysics across Worlds

The distinction can be introduced and put in a well-understood context
by reference to the widely agreed contingency of physicalism.

Suppose that physicalism is true of our world and that we have some
way of delineating a preferred set of ingredients, the physical ingredients,
to be the building blocks for everything in our world. We can think of
these ingredients as those that appear in some improvement of our current
physical sciences: physics, chemistry, and biology. What is important for what follows is that there exists a preferred set, not the adequacy of one or another delineation of the preferred set. This means we can be casual about how the delineation job gets done, provided we agree that the job is there to be done—and if we didn’t agree about the latter, we wouldn’t be giving the topic of fundamental metaphysics and physicalism the time of day.

Call the preferred set ‘core-physical’ or ‘physical’ rather than simply ‘physical’ as it will be useful to reserve ‘physical’ from now on for a wider notion to be explained shortly. Among the physical will be shape, mass, force, being at so and so a distance apart, being an electron, and the like, perhaps understood in ways that improve on our current understandings. Physicalism about our world is then the doctrine that our world is a (vast) aggregation of the physical, and that’s all. Inflation, pain, consciousness, death, belief in God, the El Niño effect, the Andes, and so on come along automatically with that aggregation.

This is a doctrine in fundamental metaphysics, but it is very much a doctrine about the nature of our world. But if we take the usual view that physicalism is a contingent truth (or better that if it is true, it can be only contingently so, but for simplicity we will speak from now on as true believers in physicalism), and, moreover, that some of the worlds where it is false are worlds with, for instance, pain and inflation, we have a further question, namely: what is being said about the nature of pain and inflation, the properties, the similarities in logical space? Suppose that $W_p$ is a world where physicalism is true, our world let’s say, and $W_q$ is a world where physicalism is false, a world made of quite different kinds of stuff from ours, but a world like ours in having pain and inflation. The second question is: wherein lies the similarity between pain in $W_p$ and pain in $W_q$?

This is clearly a question in metaphysics about the (transworld) nature of pain, and it is what I have in mind by the metaphysical question across worlds.

(Some will object to the way I have used physicalism to introduce the distinction between metaphysics at a world and metaphysics across worlds. It is one thing to hold that there are worlds where physicalism is false; another to hold that there are worlds where physicalism is false which contain pain and inflation. The objectors may argue that we should all accept the first but insist that worlds where physicalism is false can at best have fool’s-pain or twin-pain, and fool’s-inflation or twin-inflation. I think this is a mistake: pain and inflation are not natural kinds, but in any case we could have made the key point in the very terms of the objection by intro-
ducing ‘core pain’, where core pain is the similarity between pain and fool’s-pain, and pointing out that the nature of core pain is a question in metaphysics.)

Although we have two questions, the answers are interconnected. If our world, $W_p$, is an aggregation of the physical, that which unifies pain in $W_p$ and $W_q$ must be a property possessed both by that which is an aggregation of the physical, and that which is not. The answer to the cross-world question must be something possessed equally in both sorts of world. This of course is something that functionalists about mental properties have been saying more or less from the beginning, but our discussion to come will be more general in flavor. We will not be presuming functionalism.

6 Aggregation and Similarity

Physicalists are parsimonious in the properties they countenance, or at least in the instantiated properties they countenance. This is why they are as against dual attribute theories of mind as they against dual substance theories (see Smart 1959; Armstrong 1968; Lewis 1966 [in 1983]). All the same, it would be quite wrong to say that physicalism holds that the only properties possessed by the items to be found in our world are physical ones, and not just for the reason that whenever something has property $P$, it has property $P \lor Q$. The reason it would be a mistake is that aggregation enlarges the stock of properties possessed (and in interesting ways, i.e., not merely by addition). Consider a world composed solely of many stationary point masses with locations in space-time fully specified by four numbers. It would be a mistake to hold that the only properties instantiated in such a world are properties like being of mass three grams, and being at <12, 7, 9, 45>. There will be aggregations that have properties like: making up a rectangular array, being smaller than the nearest triangular array, being near the center of mass of the world, and so on.6

How then should we state physicalism so as to retain its austerity about what properties are instantiated without turning it into a doctrine we should reject straight off? I think we should think of physicalism as holding that our world is nothing more than a huge aggregation of items having only physical properties, and that all the properties instantiated in our world are those you can get by mere aggregation of items with physical properties alone. This way into the question is very much in line with the idea that physicalism is a doctrine about what you (‘God’) can make using a certain stock of ingredients, with the ingredients dubbed ‘physical’. Likewise, compatibilism about free will is the doctrine that the right
aggregation of items in a deterministic world makes up subjects that sometimes act freely.

If we use ‘physical property’ to mean any property you can get from the physical by mere aggregation, we can state physicalism as the view that our world is entirely physical in nature. This gives the desired kind of supervenience result. If our world is an aggregation of the physical, and nothing more, any minimal physical duplicate of our world will be an aggregation of the physical, and nothing more. The worlds will in consequence agree in all physical respects, that is, in all respects if physicalism is true.

What this means is that the only properties physicalists should countenance are the similarities that come from conjoining. Our world, for them, is one physical, thing plus another physical, thing, plus yet another physical, thing, plus . . . Conjoining delivers new similarities between aggregations, ones that outrun the physical, similarities, and these similarities, old and new, exhaust those to be found in our world. Can we state a closure principle for these similarities? One thing we can say is that the new similarities are closed under necessitation. That follows from supervenience. The interesting question is whether or not they are closed under a priori determination. We will look at this question in the next section as it arises in general, not just as it arises for the case of physicalism.

7 A Priori Closure and Similarities under Aggregation

Two houses that are made up of exactly the same kinds of items arranged in exactly the same kinds of ways are alike and unlike in exactly the same ways to each other and to anything else. Likenesses and unlikenesses supervene on composition, in the strong sense of composition that includes mode of composition along with constituents. Of course within a given world two houses with the same composition may differ through having different locations and surroundings, but for any two worlds differing at most in having house H1 replaced by house H2, where H1 and H2 are composed alike, H1 and H2 stand in the very same relations of likeness and difference to everything in their respective worlds. This is not a contingent claim. It is a matter of necessity. It is also a priori. What is more difficult is whether or not any given respect of likeness or difference is a priori available, given full knowledge of how the houses are composed. That the similarities are the same is a priori; what is trickier is whether or not, for any given similarity, it is a priori determined that it obtains. That question, or its generalization to aggregations, is the topic of this section.
We can say straight off that there will be similarities that are opaque. Some will be opaque in the sense of being hard for us to spot. Honeybees give information to each other about where nectar is to be found by carrying out complex bee dances in front of their hives. Finding the key similarities that carry the information is hard. Although no one doubted that the similarities were there to be found, it wasn’t easy to find them, and there was for a time, I understand, some controversy about whether or not Karl von Frisch had in fact found them (see von Frisch 1947). Some similarities will be opaque in the sense that although we can recognize them easily, it is hard to specify them. An example is that which unites all examples of a handwritten ‘a’. We humans spot the pattern easily: acquiring the relevant recognitional capacity is part of learning to read handwriting. But it is a major task for programmers to find the relevant specification, which is why it is hard to program computers to turn handwriting into displays on monitors. We know that there is a unifying similarity; we know that we are able to recognize it; we know some things about it—for example, that the color and temperature of the token letters is neither here nor there, whereas their shapes are crucial—but we do not know the specification of the properties of the tokens that unites the token ‘a’s. We have knowledge how, but only limited knowledge that.

However, in both these cases, the relevant similarity is a priori determined by the composition of the dance and the tokens, respectively. The task of deciphering the dance of the honeybees is taken by researchers to be nothing more or less than finding the property of the way the dances are put together that is the similarity, and everything is a priori identical with itself. The same goes for handwritten ‘a’s.

The thesis of a priori kind closure under aggregation is the thesis that the similarities consequent on an aggregation of elements are a priori determined by the nature of the elements and their arrangement. I wish I could think of a decisive argument for a priori kind closure. I can only say that I know of no counter-examples to it, and that when scientists seek similarities in complexes, they implicitly take it for granted. How so? What they take for granted is that there is a way of giving the similarity that is explicitly in terms of the nature of the bits and their arrangement. That is the sense in which they take for granted a priori kind closure (for more on this, see Jackson 2006).

It might be thought the water–H₂O example is a counterexample to the principle of a priori kind closure. Certain aggregations of atoms make up H₂O and are alike in being water, but this likeness is a posteriori determined, not a priori determined. But this would be a mistake. H₂O, the kind,
is a priori determined by the aggregation of elements that make up H₂O molecules, because being H₂O is a priori determined by the aggregation of those elements. This means that water, being the very same kind, is a priori determined by the aggregation of those elements. That follows from Leibniz's law. The water–H₂O example is quite different from the example of laws on views that hold that laws are necessary a posteriori. On these views we have a posteriori necessary connections between distinct properties—being light and being a first signal, as it might be—precisely what we don’t have in the water–H₂O example.

It has been suggested to me in discussion that cases where patterns become salient when items are viewed in a wider context are counterexamples to a priori kind closure under aggregation. The first five numbers in a sequence can have a commonality that only becomes salient when those numbers are viewed as part of a much longer sequence. But this is of course a misunderstanding. Salience is not to the point; what is to the point is whether each pattern, be it salient or not, is a priori determined. Equally, cases where context creates the pattern are beside the point: they are cases where the relevant level of aggregation includes the contexts. For example, one way for things to be alike is to share a common relationship to their environments; the relevant aggregations in these cases are the things plus their environments.

8 Optimism about What We Can Do with Words

It is obvious that there are cases where we cannot capture some similarity or other. We are finite beings with limited sensory and computational powers. We can be sure that the complexity of our world often defeats us. Optimists about what we can do with words insist, however, that in cases where we reliably detect similarities we can in principle find—or sometimes create—words that serve to capture the similarities, where to capture the similarity is to make explicit wherein it lies. An easy case is being circular. We have the word ‘circle’ that serves to pick out closed plane figures alike in being circular. Moreover, it is easy to find words that capture the relevant feature of the elements whose aggregation makes up any given circle: namely, ‘lying in a plane, equidistant from a given point’. We have, that is, an illuminating answer to the question, What is it about the bits of a circle that make it the case that they are bits of a circle?

Other cases are much harder, but we take it for granted that the job can be done. Although, as noted above, it is hard to put into words the similar-
ity that we latch onto when we classify some letter token as an ‘a’, it is possible with enough work. We know that programmers are not wasting their time. Again, almost certainly von Frisch solved the honeybee dance, but if he didn’t, sooner or later someone else will. Neither the idea that the uniter in these kinds of case is sui generis, nor the idea that it is beyond our powers of articulation, is credible. What the optimists hold is that whenever we finite creatures can reliably detect and report on similarities among aggregations, it is within our powers to give an illuminating answer as to where the similarity lies in terms of that which is aggregated. I have no proof of optimism any more than I have a proof of a priori kind closure, but I note that researchers into bee dances, programs that transform handwriting into print, face recognition programs, writers of grammar books, and the like take it for granted all the time.

If von Frisch solved the bee dance problem, how come we philosophers have so much trouble solving the corresponding problem for being alive? We can reliably recognize and report on whether or not something is alive, and we have the word ‘alive’ to record what we recognize and to do the reporting job. It would seem that we should be able to give in words the key similarity that unites the alive. But that would seem to be no more than the task of giving an illuminating analysis of life, and that has proved to be very hard and, many would say, impossible (the example is from Block and Stalnaker 1999). However, the problem posed by the example of life is not that of verbalizing the similarities. It is not plausible that we use ‘life’ for some mysterious, sui generis similarity; perhaps we once did (or so I have been told by those who know the history of the debate over vitalism), but that is no longer the case. The problem is choosing which similarity to pick out for the honorific. When we optimists hold that the similarity can be put into words, what we mean is: for every similarity that we can report on and recognize, it is possible for us to capture it in words. We do not mean: for every descriptive term, there is a single similarity and we can capture it in words. We said earlier that descriptive terms serve the function of giving information but sometimes the information is vague and varies with context.

Let us now look at a much harder case: pains for physicalists. Pains are alike in ways we can detect in both ourselves and others. If physicalism is true, we are detecting a likeness in aggregations of the physical. We have a word for the likeness, namely, ‘pain’ (or ‘douleur’ if we speak French, ‘dolore’ if we speak Italian, etc.). But this word does not give us the similarity in terms of the nature of the aggregations. A priori kind closure tells us that there is a similarity in physical terms, and optimism
tells us that we can find and express it in physical terms. Why believe this? Consider the alternative. If physicalism is true, we can draw up a detailed account of the ways two subjects are alike in terms of their neuroscientific states, how those states interconnect with the physical environment, the way their states evolved, and so on and so forth. If physicalism is true, a list of this kind is in principle complete. To say otherwise is to say that there is an extra similarity that outruns anything giveable in physical terms, and that is to deny physicalism. Indeed, we can write a rough recipe for finding the similarity: find the pattern in the physical that underpins the exercise of our ability to recognize pain in ourselves and in others. This recipe might be modeled on the way programmers tackle the task of finding the similarity that underpins our ability to recognize handwritten ‘a’s, or grammarians tackle the problem of articulating the pattern that makes a sentence grammatical. Or we might follow the recipe David Lewis (1994) gives which appeals to Ramsey sentences. What is important here is not the path to the answer but the claim that we can find the answer, or more precisely the claim that if physicalism is true, we can find the answer. And the appeal of the claim that if physicalism is true, we can find the answer, is that we can in principle survey all the (at all) plausible candidates given physicalism. The possibly relevant commonalities that might be whatever it is we latch onto when we judge that someone is in pain are available, and one of them must be pain, modulo vagueness.

9 Why Are the Biconditionals A Priori?

The kinds of positions in fundamental metaphysics we are considering are aggregation theses, theses to the effect that our world is some huge aggregation of things of so-and-so a kind—physical, in our favored example in which to frame the discussion (or, in the example of compatibilism about free will, to the effect that a world much like ours but made of deterministic materials would be a world with free actions—but we will stick here with the first kind and indeed with our favored example). And our answer concerning the features exemplified in these worlds is that they are, first, the likenesses and unlikenesses of the items aggregated, and, second, the similarities and dissimilarities of the aggregations of the items.

The question, then, as to whether or not pain, say, exists in one of these worlds is none other than the question as to whether or not one of these features, one of these similarities, is pain. If a priori kind closure is true, this is a priori determined by the mode of composition of the relevant
aggregations. If optimism is true, we can capture the similarity in question in words for cases where we are able to reliably detect and report the similarity, and pain is such an example. Putting both points together tells us that there will be a biconditional of the form

\[ x \text{ is in pain iff } x \text{ is so and so} \]

which is a priori true, provided the world does indeed contain pain, where ‘so and so’ is the relevant capturing in words of the similarity. Mutatis mutandis for desire for tea and so on.

We noted earlier that the most attractive version of physicalism may (may) be a mild form of eliminativism. There are no pains as we folk conceive of them but there exist pain*s, where pain*s hurt. In this case what will be a priori true will rather be

\[ x \text{ is in pain* iff } x \text{ is so and so} \]

With hindsight it may well be best to read what Smart, Armstrong, and Lewis were after as not so much a priori truths of the form

\[ x \text{ is in } M \text{ iff } x \text{ is so and so} \]

but ones of the form

\[ x \text{ is in } M^* \text{ iff } x \text{ is so and so} \]

where if \( M \) is belief, \( M^* \) is enough like belief to do all the worthwhile jobs we give belief, and if \( M \) is pain, \( M^* \) is hurts but maybe isn’t quite pain as we folk conceive of it, and so forth. Smart, Armstrong, and Lewis allowed themselves a degree of conceptual housekeeping, but this does not affect the point that somewhere or other they were committing themselves (rightly by our lights) to the existence of a priori true biconditionals.

Different people mean different things by the ‘Canberra Plan’, and, as I say right at the beginning, the search for these kinds of biconditionals has a very long history that has nothing especially to do with Australia’s national capital, but one thing to mean by the Canberra Plan is the search for biconditionals like the above for the usual suspects.

Acknowledgments

I have worried about the topic of this paper from the day I first heard of physicalist theories of mind. I am conscious of many debts including especially to the three people named in the first sentence, and to Michael Bradley and David Braddon-Mitchell.
Notes

1. And there will be some overlap with what I say in Jackson 2005 and Jackson 2006.

2. For the debate see, e.g., Lowe 1987, on the antitemporal parts side, and the reply in Lewis 1988 on the protemporal parts side. For more references and more discussion see, e.g., Lowe 1998.

3. See “Introduction” to Lewis 1986a, ix–xviii. His ‘arrangements’ are our ‘aggregations’.

4. Kripke (1980, 153), expresses the idea that pain is not something purely physical in terms of the claim that there would be more for God to do to make pain.

5. Well perhaps what we get isn’t quite consciousness as we folk conceive of it but something near enough that allows the pains to hurt. We’ll set this complication aside; but see, e.g., Lewis 1995 and Jackson 2003.

6. Some will object that I am assuming an extravagant fusion principle in making these kinds of remarks. But with some cost in circumlocution, these kinds of remarks could be rephrased to eliminate the assumption.
5 The Argument from Revelation

Daniel Stoljar

1 Introduction

The story of Canberra, the capital of Australia, is roughly as follows. In 1901, when what is called ‘Federation’ occurred—that is, when the six colonies then occupying the territory of Australia decided to join forces and become one colony—it was naturally felt that there should be a capital city. But the rulers of the two most powerful cities, Sydney and Melbourne, could not agree which of them it was to be. (Nobody took seriously the claims of any other city.) So it was decided to build a completely new city more or less midway between them. In short, Canberra is constitutively connected to compromise.

In philosophy, what is called ‘the Canberra Plan’ is constitutively connected to compromise too; at any rate, the sort of philosophical project for which people use this term often or always involves articulating a compromise or replacement conception of some of the central notions both of philosophy and of ordinary life. The reasons for the compromise usually start from a commitment to a general metaphysical thesis about what the world is like, namely, physicalism. According to physicalism, the world is (in some hard-to-define sense) fundamentally or basically physical. We are then asked to agree that physicalism is inconsistent with various intuitively plausible claims about the nature of apparently existing things, such as people, colors, values, free will, experiences, ordinary physical objects, causation, and so on. How is this inconsistency to be resolved? The solution offered by proponents of the Canberra Plan is to compromise: to spell out replacement conceptions which on the one hand may reasonably be interpreted as successor conceptions to the intuitive or ordinary conceptions that cause the problem, but which on the other are compatible with the truth of physicalism. The guiding idea is that, though the
intuitively plausible claims about people, colors, and the rest partially constitute our ordinary conception of the natures of these things, there is no reason to treat them as nonnegotiable or sacrosanct.

This sort of philosophical project is vulnerable to the criticism that the search for replacement notions is unmotivated. If there is no conflict between physicalism and our ordinary conceptions, or if the reasons for supposing there is a conflict turn out to be without foundation, there is no stimulus to articulate replacement notions in the first place. One way to develop this criticism, a way defended in various places by Huw Price (e.g., Price 1997, 2004), is to say that the arguments for the inconsistency presuppose a unitary view about the function of assertion. In fact it was Huw Price and John Hawthorne (see O’Leary-Hawthorne and Price 1996) who originally coined the term ‘Canberra Plan’. Their idea was that just as Canberra is a place that does not contain or appreciate the diversity of other Australian cities, so too the Canberra Plan is an approach to philosophy that does not appreciate the diversity of the functions of assertion. Apart from the dubious sociological comparison on which it is based, this criticism suffers from being overly ambitious. What it entails is that problems about (say) experience and (say) causation have a similar intellectual origin, that is, not attending to the different functions of assertion. However, though this might be true, there is so far as I can see no compelling reason to believe it is true. Why should philosophical problems have a common origin rather than a whole series of unrelated origins? After all, scientific problems don’t in any serious sense have a common origin—why should philosophical problems?

A less ambitious way to implement the criticism that the Canberra Plan is unmotivated is to proceed case by case; that is, to consider in detail the arguments for the inconsistency in particular instances and to see whether those arguments are sound. This paper is an exercise in this less ambitious strategy. I will focus on the Canberra Plan as it emerges in one part of David Lewis’s discussion of the notion of experience. In his 1995 paper “Should a Materialist Believe in Qualia?,” Lewis suggests that the notion of experience that is implicit in folk psychology—that is, our ordinary notion of experience—is governed by a principle which he calls there ‘the Identification Thesis’ and elsewhere calls—following Mark Johnston 1992—the thesis of ‘revelation’. Lewis goes on to argue that physicalism—or ‘materialism’ as he prefers to call it (I myself will use these terms interchangeably here)—is inconsistent with the existence of experiences so conceived, and he then goes on to articulate, in accordance with the Canberra Plan, a replacement conception of experience. In what follows,
I will agree with Lewis that physicalism is inconsistent with the existence of experience so conceived but will raise some questions about whether experience ought to be conceived in this way. I will also argue, more generally, that there is no reason to believe that revelation is true.

My plan is this: In sections 2 to 7 I will set out what I take the doctrine of revelation to be. In sections 8 and 9 I will set out the argument from revelation. In sections 10 to 14 I will discuss Lewis’s response to this argument, and related responses. I will close in section 15 by returning to the Canberra Plan and asking what, if anything, our ruminations have revealed about it.

2 The Basic Idea

According to the thesis of revelation, having an experience puts you in a remarkable epistemic position. You know or are in a position to know the essence or nature of the experience; the only thing left to learn are facts about the experience that are nonessential or accidental—that is, facts extraneous to its essence. Suppose, for example, I have an itch in my toe. If revelation is true, I know or am in a position to know the essence or nature of the itch. The only thing left to learn are various facts about the distribution of something that has this essence or nature; for example, whether other people have similar itches, and when, and what having such itches causes and is caused by. Does the dean have a similar itch? Did he have it last Wednesday during the faculty meeting? Did it last the whole meeting? Of course questions of this sort raise further philosophical issues. For example, skeptics about other minds will say that I cannot know or justifiably believe that the dean has a similar itch. But skepticism about other minds is not to the point when it comes to revelation. Revelation says only that having an experience puts you in a position such that you have only the accidental truths about the experience left to learn. Whether you will learn those truths in fact is irrelevant.

This epistemic position—the one that, according to revelation, I am in with respect to this itch if I have it—is to be contrasted with the epistemic position I am typically in with respect to other things. Suppose, for example, I have a diamond in my pocket. It does not begin to follow that I know the essence of diamonds, or that the only things left to know are various facts about the distribution of things with this essence or nature—whether other people have diamonds in their pockets, and so on. For one thing, merely having a diamond in my pocket puts me in no epistemic position whatsoever. I may have a diamond in my pocket and have no idea what
diamonds are. Moreover, even if I know what diamonds are, in the ordinary sense that most of us do—I am able to reliably pick out diamonds at the jewelers, I know that the biggest diamonds come from South Africa, and so on—I may be quite ignorant of the chemical nature of diamonds, and so ignorant of something that is surely essential to them. So, according to proponents of revelation, having an itch puts one in an epistemic position with respect to itches that having a diamond does not put one in with respect to diamonds. I am in a position with respect to itches (and experiences more generally) such that the only thing left to learn are accidental truths. But I am not in that position with respect to diamonds.

Revelation has been held to be plausible both about experiences themselves, and for properties that experiences are by reputation intimately connected to, such as colors. In fact it is the literature on color that contains the largest discussion of revelation (cf. Johnston 1992; Strawson 1988, Lewis 1999, chapters 19 and 20; Jackson 1998a; as noted above, the terminology is due to Johnston). In the version that pertains to color, the thesis says that having an experience of, for example, red puts you in a remarkable epistemic position with respect, not to the experience of red, but to red itself. In particular, you know the essence or nature of red. The only things left to learn are accidental features. Whether revelation is true for colors in addition to experiences of color, and to experiences in general, is an interesting issue, and raises some questions that do not arise in the experience case. But, as in Lewis 1995, “Should a Materialist Believe in Qualia?,” I will concentrate here on experiences rather than on what experiences are of.

3 Revelation and Other Epistemic Principles

To say that revelation is true of experiences is in effect to advance a principle about the relation between experience, on the one hand, and knowledge or justified belief, on the other. For it is to say that if you have an experience, then you are in some sort of epistemic position with respect to it. But there is a whole class of different principles along these lines (see Alston 1971), and revelation is a fairly extreme principle within that class. It will be helpful in what follows to contrast revelation with two other such principles.

The first principle about the relation between experience and knowledge or justified belief that is not revelation is (what I will call) self-presentation. According to self-presentation, having an experience puts you in a position to know or justifiably believe that you are having the experience. Self-
presentation is neither necessary nor sufficient for revelation. It is not sufficient, for it is possible to be in a position to know that a property is instantiated without being in a position to know the essence of that property. Suppose I know perfectly well that there is a diamond in my pocket, and so know that the property of being a diamond is instantiated in my pocket. If I know there is a diamond in my pocket I am in a position to know it: trivially, if you know something you are in a position to know it. Am I then in a position to know the essence of diamonds? Surely not—knowing the essence of diamonds would at least involve knowing their chemical composition and maybe a good deal more besides. But chemical ignoramus that I am—or anyway may be assumed to be—I know nothing of such things. All I know is that I have a diamond in my pocket. So I am in a position to know there is a diamond in my pocket but not in a position to know the essence of diamonds.

Self-presentation is not necessary for revelation, because knowledge of the essence of a property does not entail knowledge that you yourself instantiate it. Suppose that revelation is true and that on having an experience I do indeed know its essence. To know the essence of the experience is not to know its accidental features. So it is consistent with revelation that I fail to know every accidental feature of the experience. However, that I myself have the experience is surely an accidental feature of the experience; it is no part of the essence of any experience that I have it. But then it is possible that I know the essence of the experience and not know that I myself have the experience.

The second principle about the relation between experience and knowledge or justified belief that is not revelation is—what I will call—understanding. According to understanding, if one has an experience at a certain time, then one understands what that experience is at that time. To say that one understands what the experience is is not to say that one has a verbal description to hand of the experience, or that one’s understanding will persist for any period of time. One’s understanding may be fleeting, and may be in an important sense nonconceptual or subverbal. What it means rather is that one has, perhaps tacitly, a certain kind of knowledge of what the thing is. From this point of view, what understanding says is that if a creature is psychologically complex enough to undergo a certain experience at a certain time, then the creature is psychologically complex enough to know what that experience is at least in the ordinary sense that is sufficient for understanding.

What is the relation between understanding so conceived and revelation? Understanding is necessary for revelation, for to say that one knows
the essence of experience is to say at least that one understands what experiences are. But it is not sufficient, for it is as yet an open possibility that one may understand an experience and yet not know its essence. Presumably I can understand what diamonds are without knowing their essence; maybe what is true for diamonds is true for experiences. So it does not immediately follow from the fact that one understands what itches are that one understands the essence of itches. Of course proponents of revelation are in effect asserting that this does follow, and that to understand an experience is to know its essence—but this is a controversial philosophical claim, something that requires argument and not assertion.

4 Revelation and Knowing What

The connection between understanding and knowing what that I have just made is reflected in an important aspect of Lewis’s discussion of revelation: his emphasis on the connection between revelation and knowing what. Lewis says that according to revelation: “we know exactly what . . . [our experiences] . . . are—and that in an uncommonly demanding and literal sense of ‘knowing what.’ If I have an experience with quale Q, I know that I am having an experience with quale Q and will afterwards remember (unless I happen to forget) that on that occasion I had an experience with quale Q” (Lewis 1995 [in 1999], 327). For Lewis, then, knowing the essence of something is an instance of a more general phenomenon, knowing what something is.

Now, as we have just seen, understanding something is also an instance of this general phenomenon, knowing what something is. What then is the difference between understanding and revelation? Lewis says that knowing the essence of something is knowing what it is in an “uncommonly demanding sense,” and goes on to explain what he means by saying that if one knows the essence of a thing in this demanding sense, one knows that experiences have F, where F is the essence of the experience. The suggestion here is that, though there are demanding senses of ‘knowing what’, there are also less demanding senses; moreover, sometimes when we speak of someone’s understanding something we mean only that he or she knows what it is in this less demanding sense. Presumably I may know what diamonds are but not know what they are in this demanding sense, namely because I do not know their essence. So it does not immediately follow from the fact that one knows what itches are that one knows what they are in the demanding sense that Lewis has in mind. Of course proponents of revelation are in effect asserting that this does follow, and that
knowing what experiences are is to know them in this demanding sense—but again this is controversial philosophical claim, something that requires argument and not assertion.

5 Varieties of Revelation

We have noted that revelation is to be contrasted with other principles about the relation between experience and knowledge, and pointed to the connection between revelation and knowing what. We should also note that there is a good deal of variation in the way in which the basic idea of revelation might be developed. I will mention five such ways briefly, before going on to develop the last of these in more detail in the next two sections. (We will return to some of the other ways in later sections.)

(i) Revelation says that if have an experience I know or am in a position to know the essence of the experience. But which is it—know, or in a position to know? On the one hand, it is tempting to uniformly adopt the ‘in a position to know’ formulation. Revelation so interpreted does not assert outright that if I have an itch I know its essence; all it asserts is that if I have an itch then I am such that, were various conditions met—for example, were I to think harder or be smarter—I would know the essence of the experience. On the other hand, to interpret revelation this way is also to make it more obscure than it might otherwise be; after all, what exactly are the conditions such that were they met, I would know? In the presentation to follow, I try to ignore this issue as much as possible. So far as I can see, though the ‘in a position to know’ formulation is in some ways less controversial, it is nevertheless controversial enough, and the points I will raise will retain their force whichever formulation is in play.

(ii) Revelation says that if have an experience I know or am in a position to know the essence of the experience. But the knowledge at issue here could be either tacit or explicit. That is, revelation might be
the thesis that if I have an experience then the essence of the experience is in some hard-to-define sense ‘before the mind’. Or it might be the thesis that if I have the itch, then I know its essence, even if the fact that I know it is in some sense obscure to me. In what follows, I will adopt the second formulation. Surely it is implausible that those who are itchy have the essence of itchiness in the forefront of their minds.

(iv) Revelation says that having an experience puts you in a position to know the essence of the experience. But—you might think I have taken too long to get to this question—what exactly is the essence of experience? More generally, what is the essence of a thing or property? Obviously, this is one of the more central and controversial notions in philosophy. In what follows, therefore, I will simply follow Lewis in supposing that the essence of a thing is “a property of it such that necessarily it has it and nothing else does” (Lewis 1999, 328). More generally, $F$ is the essence of $a$ just in case necessarily for all $x$, $x = a$ iff $x$ is $F$. From this point of view, when the proponent of revelation says that if I have an experience I know the essence of the experience he or she implies that if I have an experience of type $E$, then I know that $E$ is $F$, where $F$ is a property of $E$ which necessarily it has and nothing else does. To know that $E$ has $F$ in this sense is to know a property of it that identifies it—it is for this reason, as I understand matters, Lewis refers to revelation as ‘the Identification Thesis’ (see, e.g., Lewis 1999, 328). (Lewis’s proposal about the essence of essence is controversial—in particular, see Fine 1994—but I will set these issues aside here. For some further but still too brief discussion, see chapter 11 of Stoljar 2006, and section 13 below.)

(v) Revelation says that if I have an experience I know or am in a position to know the essence of the experience. This looks like a thesis about experience; certainly I have been interpreting revelation so far as if it is a thesis about experience. However, there is a slightly different way to look at the matter, according to which revelation is primarily a thesis, not about experience, but about understanding what an experience is. In the next two sections, I will spell out in more detail this dimension of variation in the basic idea of revelation.

6 Revelation and the Part–Whole Relation

As a preliminary to discussing the relation between revelation construed as a thesis about understanding and revelation construed as a thesis about experience, it is helpful to consider a comment made by Lewis that is initially extremely strange.

Lewis rejects revelation as it pertains both to the experience and to color, but he also says: “Maybe revelation is true in other cases—as it might be
for the part–whole relation” (“Naming the Colors,” chapter 20 in Lewis 1999, 353, n.21). This is unexpected. As we have noted, revelation is normally thought to apply to experiences or perhaps to properties that are by reputation intimately connected to experiences, such as colors. How then could revelation be true of the part–whole relation? We don’t have experiences as of the part–whole relation.

The way forward here is to notice that there are two slightly different theses doing business under the label ‘revelation’. The first, which I will call e-revelation, is a thesis about what happens when you have an experience. This is, or, for our purposes, may be assumed to be:

(1) If Jones has an experience of type $E$, then Jones knows or is in a position to know, the essence of $E$.

Obviously (1) is the thesis with which we have been operating so far. But e-revelation needs to be set apart from a related thesis, which I will call u-revelation. This concerns what happens, not when you have an experience, but rather when you understand what an experience is. This is, or, for our purposes, may be assumed to be:

(2) If Jones understands what an experience of type $E$, then, Jones knows or is in a position to know the essence of $E$.

Obviously (2) is different from (1). To say that experience puts one in an epistemic position is not the same as saying that understanding does. More generally, e-revelation is not u-revelation.

This distinction permits us to explain Lewis’s initially odd remark about whole and part. When Lewis says that revelation is or might be true of the relation of whole and part, he does not mean to be invoking e-revelation, and therefore claiming that having an experience as of the relation of whole and part puts you in a position to know the essence of that relation. This claim lacks clear sense, because it is unclear what it is to have an experience as of the whole and part relation. Rather he is invoking u-revelation, and means that if you understand what the relation of whole and part is, you know the essence of the relation. More generally, since u-revelation is a thesis about understanding, and is not a thesis about experience, it might be applied without oddity to domains, such as whole and part, in which experience has no role.

7 Revelation Identified

We have distinguished revelation construed as a thesis about experience—e-revelation—from revelation construed as thesis about understanding—u-
revelation. But to distinguish these theses is not to deny they are closely related. On the contrary, there is a third thesis, which when combined with (1) will generate (2). This third thesis is none other than a version of the principle of understanding that we considered above. This is, or, for our purposes, may be assumed to be:

(3) If Jones has an experience of type $E$, he thereby understands what an experience of type $E$ is.

As we have noted, like both revelation and self-presentation, understanding—that is, (3)—is one of a class of theses about experience, on the one hand, and knowledge or justified belief, on the other, but it is a fairly plausible principle within this class. So the two points we have isolated are these: first, e-revelation is distinct from u-revelation; and second, e-revelation follows from u-revelation together with a plausible further principle according to which if you have an experience of a certain type, then you understand what that experience is.

These considerations suggest a working hypothesis about the relation between e-revelation and u-revelation. The hypothesis is this: the basic doctrine at issue here is u-revelation; e-revelation is something that follows from this basic doctrine together with the view that if you have an experience you understand it. To put this differently, revelation may be factored into two claims. The first claim is understanding, namely, that if you have an experience, then you understand what it is; the second claim is u-revelation, namely, that if you understand what an experience is, you know or are in a position to know its essence. As we will see as we proceed, this working hypothesis is important when we consider the plausibility of revelation.

The fact that there are two different principles doing business under the label ‘revelation’ threatens to introduce a terminological confusion into our discussion. One way to avoid this confusion is to adopt the policy of always referring to u-revelation or e-revelation. In what follows, however, I will continue to talk without qualification about revelation. What I will mainly have in mind is e-revelation but readers should be able to discern from the context which notion is intended.

8 Revelation against Physicalism

So far I have been concentrating on what revelation is. I turn now to the question of what follows from revelation, and in particular, to the argument that if revelation is true, physicalism is false.
Physicalism can obviously be understood in a myriad of ways, but here I will concentrate on the version according to which experiences of type $E$ are identical to physical events of a certain type; more briefly, physicalism entails that $E$ is identical to Phys (where Phys is some relevant type of physical state). From this point of view, the argument from revelation proceeds by pointing out that when revelation is combined with two agreed-on facts, it is inconsistent with physicalism.

Why is it that, when revelation is combined with two agreed-on facts, it is inconsistent with physicalism? Well, consider the following four claims:

(4) If Jones has an experience of type $E$, then Jones knows or is in a position to know the essence of $E$.

(5) Jones has an experience of type $E$.

(6) The following is one essential truth about $E$: having an experience of type $E$ is identical to being in Phys.

(7) Jones does not know, and is not in a position to know, that having an experience of type $E$ is identical to being in Phys.

It should be clear, first, that these four claims are inconsistent. If (4) and (5) are true, then Jones knows the essence of $E$. But if (6) and (7) are true, Jones does not know the essence of $E$, because to know the essence of $E$ would be to know that it is identical to Phys, and Jones does not know that. On the other hand, (4) follows from revelation; (5) is a statement of fact about Jones that we can assume or at least stipulate to be true—this is one of the agreed-on facts I mentioned; (6) follows from—or is an instance of something that follows from—the truth of physicalism; and (7) is another statement of fact about Jones that we can assume or at least stipulate to be true—this is the second agreed-on fact I mentioned.

If (4) to (7) are inconsistent, one of them is false. Setting aside (5) and (7), the only options are to deny (4) or (6). Friends of revelation argue that since (4) is true, (6) is false. That is the argument from revelation to the falsity of physicalism.

9 Options Revealed

How to respond? No doubt there are a number of possibilities, but three are particularly salient:
Response 1  Accept the argument as sound, and so reject physicalism on this basis.

Response 2  Reject the argument by rejecting the thesis of revelation outright. This response might be put in rather blunt fashion as follows: “Sure, if revelation is true, we are in trouble. But why believe it? Revelation is just an insane fantasy that has been foisted on ordinary thought about experience by philosophers for their own dastardly purposes. There is no reason to take it seriously.”

Response 3  Distinguish two conceptions of experience, an ordinary conception and a replacement conception. On the ordinary conception, (4) is true and (5) is false; more generally, revelation is true, and there are no experiences. On the replacement conception, (4) is false and (5) is true: there are experiences and revelation is false. This response, as I understand it, is the response of the Canberra Plan.

I will not attempt to discuss response 1 here beyond pointing out that presumably the reference to physicalism in the argument from revelation is not essential. Suppose a spiritualist identified having an experience of type \( E \) with being a thought in the mind of God. Presumably we could in that case imagine a person Jones—perhaps a member of some atheist cult—who didn’t know this, and yet had an experience of \( E \). The argument from revelation, or a counterpart argument, could then be used to refute the spiritualist hypothesis about experience just as much as the materialist hypothesis. Nor is the reference to fairly well-known metaphysical positions essential. Suppose (never mind why) someone identified having an experience with having a certain arrangement of pumpkins. The argument from revelation, or a counterpart argument, could then be used to refute the pumpkin hypothesis just as much as the materialist or spiritualist hypothesis. In short, to accept the argument from revelation is tantamount to accepting a rather uncompromising form of primitivism about experience according to which itches, for example, are primitive items in the world, wholly distinct from everything else. Of course, there are questions about the plausibility of this sort of primitivism, but I will not consider these, or option 1, further in this essay.

Turning then to the other options, it might appear that there is a distinction without a difference between responses 2 and 3. Surely a proponent of response 2 will agree that there is a conception of experience that incorporates the notion of revelation. (If there were no such conception, it would be a simple matter to invent one.) And surely too a proponent of response 2 will agree that there is a conception that does not incorporate...
the notion of revelation. To that extent then, response 2 and response 3 are on all fours: both accept that there are two conceptions of experience, one which incorporates revelation and one which does not; and both accept that if one adopts the conception that incorporates revelation then physicalism is true and there are no experiences, whereas if one adopts the conception that does not incorporate revelation, then physicalism is false and there are experiences. What then is the difference between response 2 and 3?

The difference between response 2 and 3 lies in their respective assessments of the two conceptions of experience that are at issue, the one that incorporates revelation and the one that does not. For proponents of response 2, the conception that incorporates revelation is a conception that has no claim on us. It is simply a conception that some philosophers have invented for purposes unknown. For proponents of response 3, by contrast, the conception of experience that incorporates revelation is the ordinary conception, or at least is a very natural one to adopt. Indeed, it is this claim of ordinariness or naturalness which as I understand things is central for proponents of the Canberra Plan. For them it is crucial that the conception of experience that causes the problem is our ordinary or natural conception. It is this fact that prompts the distinctive feature of their position, namely the search for a replacement conception of experience.

These considerations suggest the following strategy when thinking about how to respond to the argument from revelation, namely, see if the conception of experience that incorporates revelation is in fact one that is ordinary or natural. If it is not, or if the reasons for supposing that it is ordinary or natural turn out to be without foundation, we may conclude both that response 3 is mistaken and, correlatively, that response 2 is correct. In the next part of the essay therefore, I want to consider a number of comments made by Lewis in his discussion of these matters. As we will see, these comments provide a number of potential reasons for supposing that the conception of experience that incorporates revelation is the ordinary one. But what I will argue is that these potential reasons turn out to be less persuasive than they appear at first sight. If I am right, we are free to reject the conception of experience that incorporates revelation.

10 Revelation and Folk Psychology

In developing the point that the conception of experience that incorporates revelation is the ordinary conception, Lewis says that it is part of folk psychology:
Folk psychology says, I think, that we *identify* the qualia of our experiences. We know exactly what they are—and that in an uncommonly demanding and literal sense of “knowing what.” If I have an experience with quale $Q$, I know that I am having an experience with quale $Q$ and will afterwards remember (unless I happen to forget) that on that occasion I had an experience with quale $Q$. (Lewis 1995 [in 1999], 327)

Certainly, *if* folk psychology does say we identify the qualia of our experiences, a proponent of option 3 above would be on good ground in saying that the conception of experience that includes revelation is the ordinary one. But is it true that folk psychology says this? No doubt the question of what folk psychology says is ultimately an empirical question, somewhat like the question about whether folk physics says that medieval impetus physics is true. But even so I think it is appropriate to be fairly skeptical of this idea in the first instance.

First, as we noted earlier in our discussion there are *many* principles that connect experience on the one hand, with states of knowledge or belief on the other. Why think that revelation is the relevant one? For example, maybe all that is part of folk psychology is the doctrine of understanding, according to which if one has an experience one understands what it is.

Second, to suppose that revelation is built into folk psychology forces us to make some remarkable claims about the sort of cognitive sophistication that goes along with having simple experiences. Folk psychology says, I think, that animals on occasion feel pain. (Of course some people have denied this, but their views are rightly thought of as in violation in common sense.) And it is plausible also that folk psychology says that animals who feel pain understand what pains are—at least at the time they have them. But does folk psychology also say that animals who feel pain know or are in a position to know the essence of pain? That seems incredible. No doubt your average fox is a subject of suffering, but to suppose that he is also some sort of budding Aquinas is too much to bear.

Finally—this point is strictly ad hominem—Lewis himself seems to me to be uncharacteristically equivocal on whether revelation is built into folk psychology. Speaking in “Naming the Colors” of solutions to a problem about color which appeal to revelation, Lewis says: “But we materialists must dismiss this ‘solution’ as a useless piece of wishful thinking” (1999, 353). So although Lewis says in “Should a Materialist Believe in Qualia?” revelation is part of folk psychology, at least his language in “Naming the Colors” suggests something different.
11 Revelation and Obviousness

Why does Lewis believe that revelation is built into folk psychology? He writes in “Should a Materialist Believe in Qualia?”: “Why do I think it must be part of the folk theory of qualia? Because so many philosophers find it so very obvious. I think it seems obvious because it is built into folk psychology. Others will think it gets built into folk psychology because it is so obvious; but either way, the obviousness and the folk psychological status go together” (1999, 328). Once again, however, I think we should be skeptical of the supposed obviousness of revelation.

First, an obvious claim as I understand it is something which is either pretheoretically clear, or else it is something that is obvious in the context, that is, it is something that is presupposed by all parties to a dispute. But even if revelation were true, it is hard to see how it is obvious in either of these two senses. It is certainly not pretheoretically clear, if only because of its appeal to ideas about necessity and essence. And while there might be discussions in which the truth of revelation is presupposed, this discussion is evidently not one of those.

Second, the connection between obviousness and folk psychology depends on how one is thinking of folk psychology. In “Psychophysical and Theoretical Identifications” (1972) Lewis suggests that folk psychology may be thought of as a conjunction of platitudes. If by ‘platitude’ one means something obvious, then the obviousness of revelation is indeed built into folk psychology. But in later papers, Lewis suggests, in an explicit departure from his earlier view, a slightly different account of what folk psychology is. For example, in his 1994 “Reduction of Mind,” Lewis compares folk psychology to a body of tacit knowledge about a domain; on this view, you know folk psychology somewhat in the way that you know the syntax of your native language. If you think of folk psychology in this way, it is a bit unclear—at least it is unclear to me—that obviousness is to the point. Principles of syntax might well not be obvious and yet they might be tacitly known. In short, if one takes seriously the analogy between knowledge of folk psychology and knowledge of syntax, it is quite unclear that the supposed obviousness of revelation should entail or suggest that it is built into folk psychology.

12 Revelation and Kripke

I have been suggesting against Lewis that revelation is neither built into folk psychology nor obvious. But Lewis also offers some evidence of other
philosophers who think of it as philosophically central. He writes: “Kripke seems to be relying on the thesis in *Naming and Necessity* when he writes that ‘pain is picked out by its immediate phenomenological quality’” (Lewis 1995 [in 1999], 328, n.3).

As we have noted before, in Lewis’s terminology, ‘the Identification Thesis’ simply is revelation. So what he is saying here is that Kripke seems to be relying on revelation. But relying on revelation for what purpose?

At this point, it is a little hard to interpret exactly what Lewis has in mind. One possibility is that he thinks that Kripke is relying on revelation when he says that ‘pain’ is a rigid designator; as is well known there is a dispute between Kripke and Lewis on this point. In fact, however, it is hard to see why revelation would play a role here. The question of whether a term is a rigid designator is an empirical question the answer to which follows from, or is made plausible by, various semantic judgments about how the term functions in various modal contexts. It is not clear that revelation plays an essential role in these judgments.

A second possibility is that—according to Lewis—Kripke is relying on revelation when he distinguishes heat from pain; indeed Kripke is doing just this in the passage that Lewis quotes. Here is the Kripke passage in full:

In the case of the identity of heat with molecular motion, the important consideration was that although ‘heat’ is a rigid designator, the reference of that designator was determined by an accidental property of the referent, namely the property of producing in us the sensation $S$. It is thus possible that a phenomenon should have been rigidly designated in the same way as the phenomenon as heat, with its reference also picked out by means of the sensation $S$, without that phenomenon being heat and therefore its being molecular motion. Pain, on the other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality. Thus, pain, unlike heat, is not only rigidly designated by ‘pain’ but the reference of the designator is determined by an essential property of the referent. (Kripke 1980, 152)

As I understand him, Kripke is here saying something like this. If we were to introduce or fix the reference of ‘heat’ we would wind up saying something like “let ‘heat’ denote the cause of heat-sensations.” In this case, the expression in terms of which we fix the reference is a definite description that denotes or expresses a contingent property of heat. But nothing parallel seems to be available in the case of ‘pain’. Indeed, it seems natural in the case of pain to suppose—though this is not explicit in the passage above—that, if we were to fix the reference of ‘pain’ we would wind up
saying something like ‘let pain be this feeling’. In this case the expression in terms of which we fix the reference of pain is an indexical that denotes on this occasion pain itself, and so denotes an essential property of pain.

However, if this is the right interpretation of what Kripke is saying in this passage, it is difficult to see him as relying on revelation, and for two reasons. First, Kripke is not here talking about understanding at all. However one clarifies the thesis of revelation precisely, it surely has something to do with knowledge or understanding. But the passage from Kripke mentions neither. All he says is that the reference-fixing story about ‘pain’ would not involve contingent property of the referent, while the reference-fixing story about ‘heat’ would involve a contingent property of the referent. How is it supposed to follow from this that, in the case of pain, understanding just is knowledge of essence?

It might be thought that, though Kripke’s point about reference fixing by itself does not entail anything about understanding, there is a natural extension of what he says that does. According to this extension, to understand heat is to know that it is the cause of heat sensations, whereas to understand pain is to know that it is this feeling. In short, understanding heat turns is interestingly different from understanding pain, for to understand heat means knowing a contingent fact about heat, that is, that it is the cause of heat sensations, whereas understanding pain does not involve knowing a contingent fact about pain.

However—and this is my second reason for supposing that Kripke is not relying on revelation—though it is true that one might develop what he says in this way, this fact provides no support for the idea that Kripke is presupposing revelation. For, if this idea about what is going on is correct, what Kripke is saying is that understanding pain means knowing that pain is this type of experience. And this in turn, as I understand matters, means something like this: to understanding pain is to have a certain kind of de re knowledge, that is, it is to know de re of some type of experience that it is pain. But on that interpretation Kripke is evidently not relying on revelation. The reason is that (where $F$ is the essence of pain) revelation is not the doctrine that if one has an experience one knows of $F$, that pain is it. It is rather the doctrine that, if one has an experience, one knows that pain is $F$. That is, revelation does not involve de re knowledge of a property that is in fact the essence of the experience. It involves de dicto knowledge that one’s experience has that property.

Actually, this point is made by Lewis himself. Even if one denies revelation, he writes,
there is no reason to deny that the broad, *de re* content of my knowledge does in the strongest sense identify qualia. Hitherto, I have been denying that the narrow *de se* and *de dicto* content of my knowledge identifies the qualia. But the broad content is constituted partly by my narrow *de se* self-ascriptions involving acquaintance, partly by the identity of the objects of acquaintance. Thus I may know *de re* of Fred that he is a burglar, without in any sense identifying Fred. Likewise, I may know *de re* of a certain physical property that it is among the qualia of my experience, without identifying the property in question. (Lewis 1995 [in 1999], 329–330)

As I understand Kripke in the passage above, he is saying that understanding pain means knowing *de re* of a certain property that it is among the qualia of my experience. But that is consistent with understanding pain without identifying it, and so consistent with the falsity of revelation.

### 13 Revelation and Two-dimensionalism

The passage from Kripke to which Lewis draws attention emphasizes a *difference* between heat and pain; and the problem for Lewis’s suggestion that Kripke is relying on revelation in this passage is that a difference between heat and pain does not entail or suggest a *similarity* between pain and whole and part. However, there is an argument in the vicinity that might be thought to show directly that pain is similar to the relation of whole and part. Moreover, this argument exploits something that is often thought to be present at least in embryonic form in the aspects of Kripke’s discussion we just considered, namely the two-dimensional approach to meaning and understanding.

We might spell out the issues here by drawing a distinction between two ways of spelling out the notion of an essence, a distinction I have so far ignored. As I indicated in section 4, I am here operating with the proposal that $F$ is the essence of $a$ iff necessarily $a$ is $F$ and nothing else is. Another way to say this is to say that $F$ is the essence of $a$ iff $F$ in all possible worlds $a$ is $F$ and nothing else is. However, the phrase ‘in all possible worlds’ may notoriously be interpreted in two slightly different ways. One the first way, and the one that we have been adopting so far, the proposal comes out as: $F$ is the essence of $a$ iff in all possible worlds considered as counterfactual, $a$ is $F$ and nothing else is. Let us call the resulting notion of essence, the *c-essence* (for ‘counterfactual’). On the second way, the proposal comes out as: $F$ is the essence of $a$ iff in all possible worlds considered as actual, $a$ is $F$ and nothing else is. Let us call the resulting notion of essence *a-essence* (for actual). The distinction at issue here—between considering a possible
world as actual and considering a possible world as counterfactual—is a difficult one, but as a rough guide, we might say that being the watery stuff is the a-essence of water, whereas being H₂O is its c-essence. (The phrase ‘considering a possible world as actual’ is due to Davies and Humberstone 1982, but it has received considerable attention in the recent literature. See, e.g., Jackson 1998a; Chalmers 2004; Stalnaker 2004.)

Now, if we assume that it is legitimate to draw a distinction of this sort between two notions of essence, it is a simple matter to define two notions of revelation. According to the first, revelation says that to understand something is to know its c-essence; let us call the resulting version of revelation, \textit{c-revelation}. According to the second, revelation says that to understand something is to know its a-essence; let us call the resulting version of revelation \textit{a-revelation}. Obviously, it is c-revelation that has been the focus of this paper. What then of a-revelation? Well, here the surprising thing is that, while c-revelation is certainly a controversial doctrine, quite a lot of philosophers seem to adopt at least something like a-revelation as a quite general claim about understanding. Indeed, it seems to me that some of the philosophers who are attracted to two-dimensionalism and related ideas may be profitably interpreted as saying that, while c-revelation is almost always false, a-revelation is true when considered as a proposal about understanding. (Frank Jackson for instance, as I understand him, holds that to understand, e.g., water is to know that water is the watery stuff, i.e., “the kind common to the watery exemplars that we . . . are acquainted with” [Jackson 1998a, 49]. But as just indicated, the watery stuff is [what we here have called] the \textit{a-essence} of water, i.e., a property that, in all possible worlds considered as actual, water has and nothing else does. Generalizing from this example, Jackson’s position appears to be \textit{a-revelation}; to understand something is to know its a-essence. Jackson talks mainly about understanding the term ‘water’ rather than understanding water itself, but I take it that what he says in formal mode has an obvious counterpart in the material mode.)

Focusing then on a-revelation, we may now formulate the first premise of an argument to the conclusion that revelation—that is, c-revelation—is true of pain. If in general, understanding is knowledge of a-essence, then presumably understanding pain is knowledge of \textit{its} a-essence. Hence:

(8) To understand pain is know that pain is \(F\), where \(F\) is in fact the a-essence of pain, that is, a property of pain which, in all possible worlds considered as actual, it has and nothing else does.
The second premise of the argument is a thesis about the a-essence and the c-essence of pain, namely that they coincide or are one and the same.

(9) The a-essence of pain just is the c-essence of pain.

From (8) and (9) we may derive the conclusion that:

(10) To understand pain is to know that pain is \( F \), where \( F \) is the c-essence of pain, that is, a property of pain which, in all possible worlds considered as counterfactual, pain has and nothing else does.

But (10) is clearly a version of revelation as we have been discussing it. In short, if (8) and (9) are true, revelation in the sense that we have been discussing follows immediately.

Since this argument is valid, its evaluation turns on the truth of the premises. We have already seen some of the motivation for (8)—but what of (9)? It is here that Kripke really comes in to play. For it is often thought that (9), or something like it, is present in *Naming and Necessity*. The reason for believing this emerges when we look more closely at what it is to consider a possible world as actual. If we let Kripke be our guide, to consider a possible world as actual is to imagine ourselves in a world that is epistemically and qualitatively identical to the actual world. So, for example, to ask what property water has uniquely in all possible worlds considered as actual, what we need to do is imagine ourselves in every world that is epistemically and qualitatively identical to the actual world, and ask what property water has uniquely in each such world. Analogously, to ask what property pain has uniquely in all possible worlds considered as actual, what we need to do is imagine ourselves in every world epistemically and qualitatively identical to the actual world, and ask what property pain has uniquely in each such a world. However, because a world epistemically and qualitatively identical to the actual world is a world in which pain exists, it seems natural to say that the property that pain has uniquely at all such worlds simply is pain. On the other hand, it seems equally true that in all possible worlds considered as counterfactual, the property that pain has uniquely at such worlds is that of being pain. Hence the c-essence of pain simply is its a-essence.

In sum, it appears that here we have an argument that is in a sense suggested by Kripke’s discussion, and yet also suggests that revelation is true. Have we then arrived at the thought that Kripke is relying on revelation in *Naming and Necessity*? Moreover, have we arrived at the thought that the conception of experience that incorporates revelation is the ordinary or natural one?
The answer, in my view, is “no.” For even if we agree that (9) is true and that Kripke accepts it, it is not at all clear that (8) is true or that Kripke accepts it. As we noted in the previous section, when Kripke says that understanding pain is quite different from understanding heat, it is natural to view him as saying that the picture of understanding that is plausible in the case of heat, whatever that is, is not plausible in the case of pain. But this suggests that even if a-revelation or something like it is true for heat, it is not true in the case of pain. In particular, it is not the case that understanding pain is knowing that pain is $F$ where ‘$F$’ expresses the a-essence of pain. Rather what is true is that understanding pain is knowing of $F$, that pain is it. In other words, understanding pain requires a certain sort of de re knowledge. But if this is true, Kripke is not relying on revelation of any sort: he is not relying on c-revelation, and he is not relying on a-revelation either.

We might put the point another way by saying that the argument I just considered would establish that Kripke is relying on revelation only if it could be shown that Kripke is relying on both (8) and (9). However, even if we assume that Kripke is relying on (9), it remains to be shown that he is relying on (8). Perhaps some of what Kripke says suggests that he thinks that a counterpart of (8) is true in other cases; for example, perhaps a claim like a-revelation is true in the case of heat. But his discussion of the difference between heat and pain suggests that he is not relying on a-revelation in the case of pain. But this suggests more generally that he is not relying on revelation.

14 Revelation and Conceivability

The issues about two-dimensionalism and the interpretation of Kripke that we have just been discussing are notoriously difficult. As both Chalmers and Stalnaker have recently emphasized (see Chalmers 2004; Stalnaker 2004), there are many different interpretations of the two-dimensional framework. I will not try to investigate this matter further here. However, it might be supposed that there is a more basic line of thought lying just beneath the technical detail. The more basic line of thought connects the issue about revelation that we have been discussing with the much-discussed issue of the conceivability argument against materialism.

As is well known, the conceivability argument proceeds from two premises: first, that it is conceivable that there is a world identical to the actual world in all physical respects but different from it in some experiential or
The conclusion of the argument is that physicalism is false, for physicalism—at least as we have been understanding it—entails that any world identical to the actual world in all physical respects is identical to it in all respects whatever, including phenomenal respects.

Now there are obviously a host of questions to be raised about this argument, many of which have been taken up in recent philosophy. But the line of thought I want to concentrate on begins from the observation that in advancing the conceivability argument, its proponent presumably presupposes some conception of experience or other. Which conception? Well, on the one hand, it seems reasonable to suppose that the proponent of the conceivability argument is presupposing an ordinary or natural conception. The reason for this is roughly that, whatever else is true of the conceivability argument, it is at least a fairly natural argument and one which has very widespread appeal. On the other hand, one might also think that the proponent of the conceivability argument presupposes a conception of experience that incorporates revelation. Putting these two points together, we arrive at the idea that the conception of experience which incorporates revelation is an ordinary or natural one.

As against this, however, I think we should be skeptical of the idea that the proponent of the conceivability argument is relying on revelation, and for two reasons. First, it is true that the proponent of the conceivability argument presupposes some conception of experience—but why suppose that he or she is presupposing revelation in particular? To advance the conceivability argument is simply to engage in a certain sort of reasoning about, or thought about, experience. But in general we can reason about something, and think about it quite deeply, without knowing its essence. (If we could not, then reasoning could never lead to knowledge of essence.) Why should it not be that the sort of reasoning which a proponent of the conceivability argument is engaged in is of this sort?

Second, if the proponent of the conceivability argument relies on revelation then the proponents of a number of other arguments would have to be relying on revelation too. For example, consider the perfect actor argument against behaviorism. As I understand it, the perfect actor argument proceeds from two premises: first, that it is conceivable that there is a world identical to the actual world in all behavioral respects but different from it in some experiential or phenomenal respect (a world of perfect actors); second, that if this is conceivable then it is possible. The conclusion of the
argument is that behaviorism is false, for behaviorism entails that any world identical to the actual world in all behavioral respects is identical to it in all phenomenal respects. Now, so far as I can see, the idea that the conceivability argument against materialism presupposes a conception of experience which incorporates revelation is no more and no less plausible that the related idea that the perfect actor argument against behaviorism presupposes that conception.

But this observation is devastating for the proponent of the Canberra Plan, and equivalently for option 3. For recall that the proponent of the Canberra Plan is suggesting that we should reject the conception of experience that incorporates revelation. If that conception is being presupposed in the perfect actor argument, then presumably that argument too should be rejected. But here we have reached bedrock: for the perfect actor argument should not be rejected. (At any rate, nobody rejects it. At any rate, it is assumed to be sound by most people who teach and think about philosophy of mind.) But if it is not to be rejected, it is not to be rejected because it presupposes a conception of experience that incorporates revelation. (For further discussion of the comparison between the conceivability argument against materialism and the perfect actor argument against behaviorism, see Stoljar 2005, 2006a,b.)

15 Conclusions

I began with some notes on a certain program in philosophy, the Canberra Plan. What I have been doing is criticizing the plan by considering one piece of philosophy in which it is instantiated: Lewis’s suggestion that, since revelation is built into folk psychology, we need to replace the ordinary conception of experience with a revised conception. My discussion has fallen naturally into three parts:

In the first part, sections 2 to 7, I set out what I take to be the thesis of revelation, emphasizing that revelation is best seen as a principle about what happens when you understand, for example, experiences, as opposed to what happens when you have experiences.

In the second part, sections 8 and 9, I distinguished two ways of responding to the argument from revelation against physicalism, and suggested that the distinction between them boils down to whether the conception of experience that incorporates revelation is in any sense a natural or central one. If it is not, then one may reject the argument as starting from a conception of experience that we have no need to respect.
In the third part, sections 10 to 14, I considered a number of reasons for supposing that the conception of experience which incorporates revelation is a natural or central one, many of which I extracted from Lewis’s discussion of these matters: the idea that it is part of folk psychology, the idea that it is obvious, the idea that is presupposed in Kripke’s discussion of heat and pain, the idea that it is part of two-dimensionalism, and the idea that it is presupposed in the conceivability argument. In my view the suggestions prove to be unpersuasive. The conclusion to be drawn therefore is that response 2 rather than response 3 is the correct response to the argument from revelation: we are free to reject the argument as being driven by a conception of experience that we have no reason to endorse. Pain may be different from heat, but that does not make it similar to the part–whole relation.

What do these points tell us, if anything, about the general status of the program of the Canberra Plan? Well, one thing they tell us is that one motivation for a particular instance of the Plan is misguided. If the source of the conflict between experience and physicalism is revelation, then in my view there is no conflict. And, if there is no conflict, there is no motivation emanating from this argument to articulate a replacement conception of experience.

However, lying behind the topic of revelation that I have been concentrating on is a second, much larger, issue about the extent to which one philosophical problem can be treated as a template for another. At the beginning of the essay, I briefly considered a suggestion of Huw Price, that the proponents of the Canberra Plan assume an overly uniform conception of assertion. Price’s idea is to be criticized, I said, on the ground that it assumed that philosophical problems are in a certain sense all of a piece. But proponents of the Canberra Plan, just as much as Price, are vulnerable to that sort of point. Both Price and the Planners assume that philosophical problems exhibit some sort of deep similarity, and that in consequence there are meaningful generalizations that can be made about the origins of these problems and how to solve them. But what is the rationale behind this assumption? Of course it is useful to compare and contrast, but the idea that there is or should be some sort of structure for philosophical problems seems to me to be quite unlikely and at any rate is has not been argued. So we might suggest that what goes wrong in the Canberra Plan is not, as Price suggests, that it assumes an overly uniform conception of assertion; what goes wrong is that it assumes an overly uniform conception of philosophical problems.
I presented previous versions of this paper to audiences at the University of Aarhus and the Australian National University. I am very much indebted to the very good comments I received at both places. For further comments and discussion, I am greatly indebted to: David Chalmers, Christopher Daly, Nic Damnjanovic, Jakob Hohwy, Adam Pautz, and Derk Pereboom.
6 Names, Plans, and Descriptions

Fred Kroon

1 Introduction

Kripke’s arguments against descriptivism as a theory about the meaning or content of proper names and natural kind terms are as famous as they are puzzling. The most famous of the arguments, the modal argument, points out that there are possible circumstances in which the entities denoted by such terms lack the descriptive properties that competent speakers or communities associate with the terms. Take the names ‘Aristotle’ and ‘gold’, and consider any of the descriptions speakers or communities might associate with the names (say, ‘the Greek philosopher who taught Alexander the Great’ and ‘the rare heavy yellow, malleable metal’). Because there are worlds in which Aristotle did not teach Alexander the Great, and worlds in which gold is not yellow or even malleable, the modal argument reached the conclusion that the names ‘Aristotle’ and ‘gold’ do not have these or—by parity of reasoning—any other descriptive meanings. Kripke called such terms ‘rigid designators’, where a rigid designator designates the same object or kind in every world in which the object or kind exists.

It didn’t take long for philosophers who still thought of themselves as descriptivists to see through the argument. For descriptions were themselves often used in modally similar ways. It was easy to discern a sense in which the Greek philosopher who taught Alexander the Great might not have taught Alexander the Great (in theoretical terms, the descriptions can be accorded wide scope), and it was similarly easy to see that in no sense was it necessary that the individual (if any) who was actually a Greek philosopher who taught Alexander the Great should have ended up teaching Alexander the Great (here using the linguistic rigidifying operator ‘actually’; see Stanley 1997). But while such observations seemed to blunt the modal argument against descriptivism, most philosophers who thought
about Kripke's attack were not persuaded by this response. After all, Kripke also wielded very powerful *epistemic* and *semantic* arguments against descriptivism: even if the modal difficulties could be settled in this way (which Kripke denied), the epistemic and semantic difficulties could not. It is not an a priori epistemic fact that Aristotle is whichever Greek philosopher taught Alexander the Great, and it is not a semantic fact that, given an arbitrary world \( w \), the individual that the name ‘Aristotle’ designates in \( w \) is whichever Greek philosopher taught Alexander the Great in \( w \). Kripke’s own thought experiments suggested instead that what the name ‘Aristotle’ designates at a world is the individual in that world to whom the speaker is connected by appropriate causal-historical chains that began with the use of the name by speakers who had more or less direct acquaintance with this individual. Similarly, gold is not a priori known to be yellow and malleable, and the word does not pick out, at each world, a rare, yellow, malleable metal. Instead, what the word ‘gold’ designates at a world is the substance instantiated by certain paradigm samples that a community of users of the word is acquainted with.

But there, of course, is the rub. For presented this way, Kripke's arguments look positively self-defeating. They begin by relying on a conception of associated properties that takes these to be salient properties that specify what for speakers is the most useful and distinctive way of identifying the objects they are referring to—a way that answers not just to their own interests but also to the interests of others (speakers invited to answer the question ‘Who or what is \( N \)?’ will have in mind a hypothetical audience who want a useful answer). By their very nature, however, it is extraordinarily unlikely that such conversationally salient identifying properties are a priori known to hold of the referent of the name in question, or provide conditions that something must satisfy to be the referent of the name at a world. Properties of this kind usually emerge only as speakers try to find out more about the objects in question, and then find some of the properties more useful (because more distinctive, answering to more, or more pressing, classificatory questions, and so on) than others. Where historical proper names are concerned, properties citing great deeds emerge as almost inevitable on this picture. Although it needed Kripke to ram the point home, a descriptivism framed in terms of such conversationally salient identifying properties was bound to fail.

But of course nothing follows from this about forms of descriptivism that do not place such constraints on ‘associated’ reference-determining properties, and where the notion of an ‘associated’ property is not tied so closely to conversational purposes. What finally gives the game away, so
it seems, are Kripke’s own thought experiments, adduced to provide a “better picture” of what determines the reference of proper names and natural kind terms, but thereby undermining the very conclusions that the thought experiments were supposed to support. For it now turns out that we do have considerable insight into what determines the reference of such terms: reflection on these thought experiments yield something like a causal-historical account of how this happens, and this suggests that so long as we do not put constraints on the class of available properties, and don’t tie the notion of ‘associated’ too closely to various conversational purposes, descriptivism is in good shape after all.

This is by now a familiar, well-publicized complaint against Kripke’s arguments; it has even spawned a form of descriptivism that takes appropriate causal-descriptive descriptions to be what determine reference (a theory commonly known as ‘causal descriptivism’ or the ‘causal description theory’; see, e.g., Chalmers 2002b; Jackson 1998b; Kroon 1987; Lewis 1994; Searle 1983). But the complaint and the causal-description theory it brought in its wake still have less than fulsome support among philosophers of language (this is to put it mildly). Many think it misses the point of the original arguments. A common countercomplaint is that the proffered causal-descriptive descriptions are difficult to articulate. They require knowing the right theory of reference, a topic that continues to be the subject of intense philosophical debate. Surely we shouldn’t expect untutored speakers to know what philosophers of language find it so difficult to know! The only honest form of descriptivism for these antidescriptivists involves associated descriptions that speakers can readily articulate, and this form of descriptivism is subject to the Kripkean arguments.

The present essay is an attempt to probe and understand these worries, and then to suggest an answer. I begin by presenting Frank Jackson’s powerful version of the complaint against Kripke’s arguments. Jackson’s version emphasizes the role of implicit knowledge, but has received widespread criticism because of that emphasis. After discussing these criticisms, especially as articulated in recent work by Scott Soames, I say what I think is missing from Jackson’s version of the complaint, and, indeed, from all current versions. The remainder of the essay is an attempt to fill the gap. It argues in effect that although Canberra Planners have been keen to find a place for a respectable form of descriptivism, what is missing from their story—and what they need if the argument for descriptivism is to go through—is an acknowledgment of something like the notion of a referential plan in the institution of referring with names.
2 Jackson’s A Priori Argument for Descriptivism

In his “Reference and Descriptivism Revisited” (Jackson 1998b), Frank Jackson gives a subtle twist to the complaint that Kripke’s advocacy of a causal-historical theory is at odds with his arguments against descriptivism. Jackson has this to say:

The critics’ writings are full of descriptions (descriptions) of possible worlds and claims about what refers, or fails to refer, to what in these possible worlds. . . . But if speakers can say what refers to what when various possible worlds are described to them, description theorists can identify the property associated in their minds with, for example, the word ‘water’: it is the disjunction of the properties that guide the speakers in each particular possible world when they say which stuff, if any, in each world counts as water. This disjunction is in their minds in the sense that they can deliver the answer for each possible world when it is described in sufficient detail, but is implicit in the sense that the pattern that brings the various disjuncts together as part of the, possibly highly complex, disjunction may be one they cannot state. This is not to say that, after reflection on their classifications in the various possible cases, perhaps aided by doing a course in the philosophy of language, they won’t be able to make a good stab at stating the pattern; something like: belonging to the kind which most of the clear, potable samples, acquaintance with which lead to the introduction of the word ‘water’ in our language, is roughly right—and if you describe a case that this formula fails to cover, you do not show that there is no pattern, but that my stab wasn’t good enough. (Jackson 1998b, 212)

Jackson’s argument here strengthens our earlier statement of complaint in the following way. Supporters of the Kripke arguments often argue that Kripke’s appeal to a causal-historical theory of reference should be set aside as irrelevant to the debate between descriptivism and other theories, since it is based on certain philosophical thought experiments about what our words do, or do not, refer to in different sorts of circumstances—thought experiments that needn’t be accessible to ordinary speakers and whose results the theorists might themselves get wrong. There is thus nothing sacrosanct, so it is argued, about Kripke’s own attempts at formulating a causal-historical theory. These attempts are to be considered defeasible, and so the causal-historical conditions that result from these attempts are not a priori known to hold of the referent and are not semantically guaranteed to pick out the referent at worlds where the name is used.

The passage from Jackson shows what is wrong with this way of arguing. To begin with, the thought experiments may be the province of philoso-
phers, but that is only because they are the ones interested in the theory of reference. The thought experiments are not philosophical per se, and if they didn’t deliver insights that ordinary speakers would also assent to, we should scarcely take them seriously (cf. Braddon-Mitchell 2004, 147ff.). Why be interested in what a philosophical expert thinks if this is at odds with what ordinary speakers would say? Second, speakers’ intuitions about what a name refers to in various possible worlds build on their implicit knowledge of what it takes for the name to refer to something. What shows that they have this knowledge is that they can work out what the name refers to when presented with various possible scenarios. But this doesn’t mean that they have explicit, articulated knowledge of the possibly disjunctive pattern of properties that guides them when they say what the name picks out in the various possible scenarios. The only time, in fact, when the epithet ‘philosophical’ becomes appropriate is when speakers begin inquiring into these properties and, “perhaps aided by doing a course in the philosophy of language,” take a stab at stating the pattern. So the defeasibility of the causal-historical theory is neither here nor there. What makes descriptivism a priori true has less to do with the articulated end-product than with the fact that whenever we try to describe the referential behavior of some alleged counterexample to descriptivism we are relying on an ability to tell what the name does, or does not, refer to in some target world, and having this ability suggests that we are implicitly guided by properties that determine the right answer.

Why hasn’t this argument—call it the *a priori argument* for descriptivism—persuaded antidescriptivists that a form of descriptivism is therefore bound to be correct, even if it is not the form they had in mind? There are a number of reasons that stand out as particularly important: (1) the “no implicit knowledge” objection, (2) the “redundancy of implicit knowledge” objection, and (3) the “semantic strangeness” objection.

(1) The “no implicit knowledge” objection argues that speakers do not typically have implicit knowledge of this kind. Scott Soames, for example, thinks that the kind of knowledge shown by theorists of reference on the basis of their reflection on various possible scenarios is reflective, philosophical knowledge, based on thought experiments. Although it may in some sense be accessible to ordinary language-users in their reflective, philosophical moments, it is not knowledge available to speakers in their largely unreflective practice with names. Soames thinks that the attribution of ‘knowledge’ of this kind to ordinary speakers is tendentious in the way Plato’s attribution of a priori knowledge of mathematics to the slave boy in the *Meno* is tendentious (Soames 2005, 299).
(2) The “redundancy of implicit knowledge” objection, advanced in different ways by Devitt and Sterelny (1999) and Soames (2002, 2005), argues that postulating implicit knowledge of causal-historical properties is in any case theoretically redundant; such knowledge is not needed by a theory of reference. Thus Devitt and Sterelny claim that if causal-historical properties are satisfied, then causal-historical chains on their own can be assigned this reference-determining role, whereas Soames thinks that the instantiation of such properties presupposes antecedently attained de re cognitive attitudes. More radically, Soames thinks that the ability to use a name with the reference it has for others in one’s community may simply be the result of the name’s being in public use, together with the standing intention that one’s use of words (not just names, but any words) conform with the linguistic conventions of one’s community; if so, it would be a mistake to think of causal-historical chains, let alone causal-historical properties, as reference-determining mechanisms.

(3) The “semantic strangeness” objection argues that relevant causal-historical properties may well be quite idiosyncratic, and so not the kind of thing that can count as providing anything like the meaning of names, whether taken as the communicated content of names (we, the audience, may know little of the causal history of the speaker’s use of a name) or as the content of names in propositional attitude attributions.

Of these three objections, I regard the first two as genuine objections to descriptivism. The last objection, on the other hand, strikes me as an objection to an unduly ambitious form of descriptivism. I agree with the substance of the objection (although not with some of the lessons that antidescriptivists have drawn from it), and the reasons for this judgment will emerge at the end of the paper. For much of the remainder of this essay, I want to focus on the first two objections.

To see what is needed to answer these objections, suppose that a similar complaint is made about Smith’s inability to describe the way to Jones’s house even though he unerringly gets there. It is hard to resist the view that in this case Smith knows the route, but can’t articulate what he knows. The attribution owes much to belief-desire psychology: we can explain Smith’s getting to Jones’s house in terms of his desire to go there and his belief—in this case, knowledge (since he is unerringly right)—that the route chosen will get him there. Of course there may be other explanations available to explain Smith’s ability. Perhaps his uncanny ability is due to some kind of robotic “homing” device, and not to knowledge that he is able to exploit. But if we really have doubts along these lines, there are (defeasible) ways to determine the truth. The obvious way is to try him
out on new destinations; if he is able to learn how to get to a destination, despite the fact that he is again inarticulate when asked for a description of the route, we will surely accept that in all these cases he has exploitable knowledge of the route, but that the knowledge remains relatively implicit.

So one way of answering Soames’s skepticism is, first, to provide a positive reason, in terms of something like belief-desire psychology, for the claim that speakers do indeed have implicit knowledge of reference-determining properties that guide them in their attempts at reference, and, second, to test such an explanation by demonstrating that speakers show exploitative knowledge of such properties in a variety of creative ways. That is precisely how I want to supplement Jackson’s a priori argument. In the next section, I develop such an antiskeptical argument, and use it to answer the “no implicit knowledge” and “redundancy of implicit knowledge” objections. In the final section, I return to the “semantic strangeness” objection. For the most part, I shall focus on the case of proper names.

3 Motivating Descriptivism

Both the causal-descriptivist approach and the causal-historical approach are silent on one crucial issue: it is clear that not all properties and chains or networks of a causal-historical kind play a reference-determining role—causal-historical on its own does not mark out anything remotely like a semantic kind. So what is special about the causal-historical properties and chains that do? I think that the answer, or at any rate a crucial part of the answer, lies in the cognitive nature of such links and properties: the way they are implicated in the acquisition and transmission of information about objects of reference.

To see this, consider ordinary proper names and familiar natural kind terms. Such terms are deemed useful to a community at least in part because its members take there to be relatively rich modes of epistemic access to the objects and kinds in question. Indeed, they make use of this cognitive availability when they decide that these objects and kinds are of a type to deserve designation by a special term in the first place. What is crucial to the manner in which ‘George W. Bush’, for example, first began to be used to refer to the person who was later to become U.S. president is not simply the causal nature of the ties between a group of early users of the name and George Bush himself, but the perceptual nature of the ties linking this group to that person—a resilient type of connection, of a
type available at other times and to others, that allows those so connected to track the child through the course of its spatiotemporal meanderings.

Similarly, ‘water’ does not denote the scattered substance water simply in virtue of a certain kind of causal relationship in which users of the term, or many of them, stand to samples of water in the actual world. It denotes that substance in virtue of the fact that the substance is cognitively available to users of the name, through perception. This in turn explains the interest humans have in the substance (most are able to judge that they are confronted by stuff that is liquid, transparent, potable, sometimes falls from clouds, and found roundabout here in lakes and rivers), and explains their interest in ways of communicating with each other about it. Perception is not just one possible causal relation among many that should interest causal theorists of reference. It is, crucially, a cognitive causal relation, and it is the cognitive, epistemic character of the relation of perception that is important for a theory of reference.

But if something like this is right, we can tell a certain story about what motivates our referential practices with names. For the above reflections make little sense unless language-users have at least implicit awareness of the cognitive character of a causal relation like perception. When language-users perceive an object or (instances of) a kind, they normally expect this perceptual relationship to be resilient, and of a type available at different times and circumstances, allowing them to add to, or to correct, their stock of information about the object or kind in question. They also expect others to be able to form their own store of information, and they expect information to be able to flow from one agent to another. If they didn’t have such expectations, and didn’t believe others had such expectations, there would be little chance of systematic linguistic commerce that had the object or kind as its continuing focus: the kind of systematic linguistic commerce that is surely characteristic of a referential practice with a name.

This yields the following idea. Our referential practices with names require that we have an expectation of what I shall call the cognitive availability of objects and kinds, that this expectation is a shared or common expectation, and that we are at least implicitly aware of this common expectation. Furthermore, this kind of expectation also underlies the way in which new names are introduced. We don’t introduce names if there is nothing that can be usefully said about the objects or kinds singled out: if they are not seen as meeting our expectations of cognitive availability. We don’t, for example, introduce a name to stand for whichever individual will have the most impact on world history in 2050, say. Such an introduc-
tion would be entirely pointless; apart from speculative predictions, and some logically inferred claims, nothing could be usefully said about this person. For the same reason, we don’t introduce a name for whichever species is the most prolific undiscovered species of animal in the Amazon rainforest.

So far I have argued that the cognitive nature of the causal-historical links involved in reference is something that agents more or less self-consciously exploit in setting up, and participating in, referential practices with names. But once we acknowledge this element of exploitation, it becomes difficult to resist a descriptivized version of a causal-historical theory. For it suggests that referrers are planners: they select their referential targets on the grounds that these not only deserve to be the subject of discourse but they can be the subject of such discourse, that the targets are cognitively available in the right sort of way. And that suggests that what fixes the reference of terms for these targets is something that builds reflectively on this cognitive availability.

One way to see this is to imagine a community S of name-users whose way of introducing and using names mimics ours, is similarly focused on the importance of the cognitive availability of objects of reference, but is entirely explicit. Suppose that members of S explicitly introduce and use names on the basis of the kind of acquaintance-invoking causal properties that describe the objects and kinds referred to by name-users from our own community (as it may be, properties like being the individual whom, on the basis of my current perceptual experience, I take to be a red-haired child sitting directly in front of me or being the individual called ‘Aristotle’ whom I take to be a famous Greek philosopher on the basis of information acquired from others using the name). If we stipulate that (semantic) reference-determination for names in S goes by way of such properties, it turns out that the right kind of theory of name-reference in S is a version of (causal) descriptivism. Now, I take it that causal-historical theorists can scarcely deny the possibility of such a theory of reference-determination, even if they deny its actuality. But note that there will be no real difference in referential behavior between community S and our own community since we are supposing that the way in which speakers in our community exploit the cognitive, acquaintance-based nature of their ties to objects and kinds is replicated in community S by an explicit reliance on such acquaintance-invoking properties.

Soames, as we know, refuses to allow that name-users in our own community have even implicit knowledge of such properties. But, pace Soames, the case is very different from explaining the success of Meno’s slave boy
in terms of his recollection of knowledge obtained before birth. First of all, the adept way in which speakers in our own community are able to identify the objects and kinds they are referring to, to all appearances on the basis of just such properties, reflects the central importance in our community of the cognitive availability of the objects of our reference: such adeptness is what one should expect in a community that wants to have referential practices focused on the acquisition and sharing of information, and such adeptness is what we find in different forms, in our own community no less than in $S$. Differently put, our ability to identify the referents of names in different possible scenarios—an ability cited by Jackson in his a priori argument for descriptivism—is itself rooted in a capacity with causal-historical properties of this kind that we share with members of $S$ in virtue of our being active and aware participants in referential practices focused on the acquisition and sharing of information. From this point of view, community $S$ differs from our own in the explicitness with which such properties fulfill their role in $S$, and not in the role itself.\(^3\)

A second point is that explicitness of this sort is surely a transient matter. It is likely that members of $S$ will take shortcuts when asked to identify which objects they are referring to. If on the basis of her use of the name ‘Aristotle’, a speaker is asked who Aristotle is, she is likely to say something like ‘the famous Greek philosopher who tutored Alexander the Great’ rather than ‘the individual named “Aristotle” whom—on the basis of information I have acquired from others using the name—I take to be a famous Greek philosopher born in Stagira who tutored Alexander the Great’; the second description contains material that is likely to be both irrelevant and misleading from the point of view of the questioner. Even so, it is the second, not the first, description, that specifies how reference is determined, which is why the speaker can consistently add: ‘Perhaps Aristotle did none of these things’. But verbal shortcuts of this kind in response to queries about the identity of objects being referred to are not only useful; they are almost inevitable, given the conversational purposes served by such queries and their responses (such queries are typically meant to elicit features that are salient relative to some context). It would not at all be surprising, therefore, if, after repeated reliance on such shortcuts, speakers begin to find it difficult to articulate on demand the real reference-determining properties of the names they use. Unless we want to see such inarticulateness as having the semantic implication—surely an implausible implication—that what was previously a descriptivist theory of reference determination has now become a causal-historical one, it seems best to insist that what previously was explicit has now become implicit.
We now have the kind of argument I promised at the end of the previous section: an argument that invokes something akin to belief-desire psychology to support the claim that language-users do indeed have implicit knowledge of reference-determining properties, contra Soames’s “no implicit knowledge” objection. But can we do better? Can we also point more directly to the sort of genuinely creative ways of referring with names that might put the matter beyond contention? I think we can, by focusing on cases where speaker’s beliefs about causal-historical links play an absolutely essential role in determining reference. Given the prominent place occupied by the phenomenon of reference borrowing in debates about the way names refer, let me focus on the phenomenon of indirect reference borrowing. Here is a Kripke-inspired example of this phenomenon. Suppose Smith calls his pet aardvark ‘Napoleon’. After hearing Smith use the name in the course of proudly cataloging his pet’s achievements, Jones remarks: ‘What an interesting name for a pet! Who is, or was, Napoleon?’ As the context makes clear, underlying Jones’s use of ‘Napoleon’, there is an episode of reference borrowing, although not an episode of the usual kind. To the extent that Jones does not intend to speak of Smith’s aardvark, the reference borrowing is indirect. At this point, the causal-historical theorist may think that she should accordingly acknowledge two modes of reference borrowing: direct and indirect, with the latter based on a causal-historical link that shows the speaker to be referring to the original bearer of the name. But it is easy to see that this can’t be right, for the context on its own does not settle the question of the reference of Jones’s use of ‘Napoleon’. As it stands, the conversation need not leave us with just a single contender. Suppose that Smith called his pet ‘Napoleon’ because that was his own first name, but that his mother—a Francophile military historian—called her son ‘Napoleon’ after Napoleon Bonaparte. If Jones simply wants to know whom Smith named his pet after (perhaps he is doing research into the way children give names), then his use of ‘Napoleon’ singles out Smith. But Jones might realize that it is likely that there was an important historical person the boy was named after, in which case the query could just as easily concern the original Napoleon. (‘What an interesting name your son and his pet have. I bet the chap they were named after, Napoleon, was someone quite important.’) And it is clear that the case could be even more complicated, with the name having an even more complicated history and speakers having causal-historical beliefs that correctly capture this history which speakers can then exploit to their own referential ends.

The speaker’s beliefs play a palpable, creative role in reference determination in this kind of case. Absent the beliefs, there would have been no such
possibilities of indirect reference borrowing, and so it seems that this kind of reference borrowing should be understood in terms of associated properties like *being the individual called 'Napoleon' whose name was deliberately recycled by the boy I am talking to, and used as the name of (the animal I take to be) his pet aardvark*: that is, in causal-descriptivist terms. Of course causal descriptivists also think that the kind of direct reference borrowing focused on by causal-historical theorists is just a limiting (albeit pervasive) case of such a more general notion of reference borrowing, and that it should therefore also be understood descriptively.

So much for the “no implicit knowledge” objection. Devitt and Sterelny, both ardent advocates of the causal-historical approach, have responded to such arguments by putting forward what I have called the “redundancy of implicit knowledge” objection. According to this objection, if there is indeed implicit knowledge of descriptions that encode such causal-historical beliefs then, since the beliefs are true, there are also the causal-historical chains themselves; but in that case the implicit knowledge is redundant, and the chains can be assigned the reference-determining role on their own (Devitt and Sterelny 1999, 61). In the same spirit, Soames claims that causal descriptivists think that de re belief expressed through the use of names reduces to the availability of broadly causal, acquaintance-invoking descriptions that can themselves be understood “only if antecedently attained de re beliefs and other cognitive attitudes are presupposed” (Soames 2005, 322). He concludes that for causal descriptivists it is therefore “immaterial whether or not the speaker understands the descriptions or sentences containing them” (ibid.), showing the semantic irrelevance of the descriptions.6

But such complaints rest on a confusion. It is true, of course, that if the speaker’s causal-historical beliefs are true then this marks out certain causal-historical chains. But this does not give them a primary reference-determining role. If it did, then the semantics of certain genuine descriptions would similarly be in thrall to the causal-historical chain theory. For take any attributively construed description that specifies an object in terms of a speaker’s causal relationship to that object: say, ‘the individual, if any, in my perceptual field wearing a monocle’, where the speaker’s perceptual evidence makes him think that there is someone fitting the description. If such a description picks out an individual, there clearly is a causal, acquaintance-based connection between the speaker and the individual, and a speaker who uses this description in a sentence to express a belief thereby expresses a de re belief about this individual. It would be wrong, however, to insist that what makes it the case that the speaker’s
Names, Plans, and Descriptions

description stands for that individual is the causal connection between the speaker and the individual so that it would be redundant to invoke semantic properties of the description to explain the way the description denotes. Such a move would be an obvious non sequitur; the semantic properties of the description are what we use to single out the role of the causal connection. Similarly, it doesn’t follow that it is “immaterial” whether or not the description or sentence is understood. Indeed, if the speaker doesn’t understand the description or the sentence, he surely wouldn’t succeed in expressing any belief at all.

Perhaps a more charitable interpretation of the kind of complaint that we find Soames as well as Devitt and Sterelny making is that there is no effective difference between using names against a background of implicit knowledge of acquaintance-invoking descriptions and using them against a background of relationships of acquaintance that speakers don’t have implicit knowledge of; hence such knowledge cannot have any significant semantic role to play. But note that this cannot be shown by simply pointing out that acquaintance-invoking descriptions invoke relationships of acquaintance. It may just be the case, after all, that the sort of acquaintance that is needed for name-using referential practices requires implicit knowledge of acquaintance-invoking descriptions. That, in fact, is precisely what my response to the earlier “no implicit knowledge” objection was designed to show; such knowledge has a special role to play in an institution of naming, since it is a consequence of a requirement of cognitive availability of the objects and kinds to which we make reference with names. So the “redundancy of implicit knowledge” objection fails in this form as well.

Let me briefly deal with a final form of the objection. As it turns out, Soames disagrees not only with causal descriptivism but also with causal-historical accounts of reference. Focusing on the ability to use a name with the reference it has for others in one’s community, he argues that there are certain counterexamples to such theories that suggest that something rather different may be needed—he thinks a type of communitarian view on which such an ability “may simply be the result of the name’s being in public use, together with the standing intention that one’s use of words conform with the linguistic conventions of one’s community” (Soames 2005, 301–302). If this is right, then neither causal-historical descriptions nor causal-historical chains have a role to play in explaining a speaker’s ability to use a name with the reference it has for others.

My response to the “no implicit knowledge” objection can be parlayed into a response to these new concerns as well. Take Soames’s counterexample first. Suppose you know of a certain region in Ireland in which
residents of different towns see to it that there is always exactly one person bearing the name ‘Patrick O’Grady’, and suppose you then enter a pub in a new town in the region, and announce ‘I am looking for Patrick O’Grady, whom I am willing to pay for an interview for my new book’. In this case “you successfully use the name to refer to the man, and say something about him—not because you have acquired the name through a causal-historical chain of reference-transmission, but because you are able to speak the language of the community in which the referent of the name has already been established” (Soames 2005, 301).

Soames is right about the way the speaker bypasses relying on others for successful reference, but we should scarcely be surprised at this ability. The ‘Napoleon’ case mentioned earlier already showed us how adept speakers can be at exploiting their causal-historical beliefs. The ‘O’Grady’ example is simply another case of this kind, with the speaker able to exploit his beliefs about the name ‘Patrick O’Grady’ to gain access to a referential practice involving the name—and this time without even hearing the name-type used by anyone. Once we keep in mind the cognitive gloss that should be placed on the role of causal notions in the theory of reference, it seems that there is nothing in this to embarrass my kind of causal descriptivist.

The ‘Napoleon’ example also exposes difficulties in Soames’s positive suggestion that a speaker’s ability to use a name with the reference it has for others in his community “may simply be the result of the name’s being in public use, together with the standing intention that one’s use of words conform with the linguistic conventions of one’s community.” For no amount of gesturing toward linguistic conventions is able to identify the referent of the name ‘Napoleon’ as used by Jones. The community to which Jones belongs uses the name for Smith and his pet aardvark as well as for Napoleon Bonaparte (and perhaps many others), with a number of these uses reflecting referential practices that Jones can easily slot into simply on the basis of hearing Smith use the name and having appropriate beliefs. Linguistic conventions simply do not have the power to explain which of these is the actual referent of Jones’s token of ‘Napoleon’. If my earlier arguments are correct, only an appeal to implicitly represented properties can do this.

4 Living with Descriptivism’s Semantic Limitations

This brings us to the third and final objection: the “semantic strangeness” objection that relevant causal-historical properties may well be quite idio-
syncratic and so not the kind of thing that can count as providing the meaning or semantic content of names. And here, I think, the critic of a broadly causal-historical version of descriptivism is on safer ground. It is no doubt a controversial matter what roles a notion of content ought to play, but what is surely not contested is that content in the semantic sense should at the very least explain the possibility of communication. After all, the notion of the semantic content of an expression is tied to the notion of the semantic content of a sentence containing the expression, that is, the proposition expressed by a sentence containing the expression, and what proposition is expressed is presumably what an audience is supposed to grasp in the interests of communication. It is clear, however, that austere causal-historical properties do not have this communicative dimension of a notion of content. To begin with, there is too much variation among the broadly causal-historical properties associated with the same name by different speakers and hearers, variation that may nonetheless be utterly irrelevant to communicative success. And even if they were properties that both speaker and audience could grasp, they don’t seem to be the right kind of properties. For according to our description theorist, they are typically austere causal-historical properties like being the actual person whom I have heard or seen called ‘Aristotle’ by the people from whom I learned the name: egocentric, metalinguistic properties that appear in no way to be part of the subject matter of what we say and communicate when we utter sentences using names. (It just seems silly, not the statement of a promise and its fulfillment, to say: “Let me tell you a little bit about myself: Aristotle was a Greek philosopher.”)

Such a theory will also find it difficult to solve various other familiar puzzles about names that older versions of descriptivism were designed to solve. It will be hard for such a theory to capture the communicated content of negative existential statements like ‘Homer doesn’t exist’, for example, since there will again be considerable variation among associated causal-historical properties (I am assuming that contingently empty names like ‘Homer’ have the same kind of associated properties as nonempty names). Even more difficult for such a theory is the problem of how to understand the role of names in propositional attitude attributions, since the agents of such attributions are not likely to have knowledge of the causal-historical properties underlying reference with such names. In ascribing to ancient Babylonians the belief that Venus was a star, for example, I am clearly not ascribing the belief that the actual object that caused my token of ‘Venus’ in such-and-such a way was a star. Ancient Babylonians knew nothing about causes of my tokens of ‘Venus’ (Soames
So the causal-historical properties appealed to in this description theory lack not only the communicative dimension of a notion of content. They also lack what classical descriptivists like Frege and Russell and their followers saw as another central dimension—the propositional attitude dimension—of the notion of content (see especially Frege 1952 and Russell 1905): such properties are poorly equipped to provide a semantic explanation of the way names behave in propositional attitude attributions. Far from being the Great Hope of the descriptivist program, if a causal version of descriptivism is the best we can hope for the program appears doomed.

Note, however, that none of these proclaimed limitations of a causal description view of reference should strike us as particularly surprising if I am right about the way referential practices involving names are rooted in the cognitive availability of the objects named. Given such a motivation, and given that the way in which objects are cognitively available is going to depend on the circumstances in which speakers and hearers find themselves, it is entirely doubtful that a causal description theory can be more than a theory of reference determination. What motivates descriptivism on this picture appears to preclude it from being a theory of the generous, multidimensional kind of semantic content that early descriptivism promised to deliver. This doesn’t mean that causal descriptivism is wrong, but only that it is wrong if its notion of content is construed as the multidimensional notion looked for by early descriptivists rather than as a notion of reference-determining content (indeed, implicit reference-determining content).

It can’t be denied that on the surface this counts as a monumental lowering of descriptivism’s sights, given the ambitions with which descriptivism first came on the philosophical scene. Still, causal descriptivism is rather more potent than this pessimistic assessment may suggest. For there are good arguments in favor of retaining a place for a causal-descriptivist kind of content even in solutions to the other problems that worried the early descriptivists—the place will just be more complex. Attempting a full-scale defense of this view would be much too ambitious, but I hope to show that it is a reasonable view for at least the first of the problems mentioned above: the fact that austere causal-historical properties do not have the communicative dimension of a notion of content. I claim that something like the communicative dimension of a notion of content can nonetheless be recovered from such properties.

With this aim in mind, consider a speaker, Jones, who utters the sentence ‘Aristotle wrote many more works than are attributed to him’. What
does Jones thereby succeed in communicating to his audience? As it stands this question is clearly far too broad, for the possibilities seem endless. If Jones’s audience knows that he is using the name to stand for the famous ancient Greek philosopher Aristotle whom they all believe to have been the tutor of Alexander the Great, then what is communicated includes the claim that the philosopher-tutor of Alexander the Great wrote many more works than are attributed to him. That, at any rate, will be one of the beliefs they acquire as a result of Jones’s utterance (assuming they believe what Jones says). Successful communication of this or that content may depend on the audience’s beliefs (Soames 2002, 72–86). But communication in this sense is not relevant if we are interested in the semantics rather than the pragmatics of communication: if we are interested in the question of what Jones’s words allow him to say simply in virtue of using those words, independently of beliefs held by Jones’s audience.

We can say more by attending to what makes it the case that Jones's token of ‘Aristotle’ stands for the philosopher Aristotle. On the causal-descriptivist picture, these are certain implicitly represented reference-determining properties that single out the philosopher. Their more or less idiosyncratic nature means that there is little prospect that hearers will grasp these properties, or will all grasp them equally well. Yet that is clearly no bar to communication, and it is easy to see why: hearers understand that Jones is like them, the recipient of information based on his acquaintance with some individual called ‘Aristotle’. If they understand enough of the context (understand in particular that there is talk of a famous Greek philosopher rather than a famous Greek shipping-magnate, say), then they understand enough to know that Jones is speaking of a particular person whom they too are acquainted with under the name ‘Aristotle’. They recognize Jones’s participation in a referential practice that is focused on this person, a practice they too can access for purposes of acquiring and sharing information. Else they decide that they have not heard of Aristotle before, in which case they will defer to Jones when they first begin thinking and talking of Aristotle. Note that they need know little else of the way Jones participates in this practice (as far as they are concerned, it may even be the case that Jones borrowed his reference for ‘Aristotle’ indirectly, by realizing that his informant had recycled the name of a famous thinker of some kind and then choosing to use the name to stand for the famous thinker). Either way, the information communicated to members of Jones’s audience is the information that—in their words, and relying on their own individual ways of reference fixing on Aristotle to put it in these words—‘Aristotle wrote many more works than are attributed to him’.
In short, Jones succeeds in communicating with his audience, and he succeeds because of the causal-descriptivist way in which the institution of naming works. Note that unlike doctrinaire Millians who think that the semantic content of a name is simply its referent, we didn’t need the notion of singular propositional content to capture such a semantic notion of communicated content.\(^{10}\) (Indeed, on the present way of understanding communicated content we should deny that two [indexical-free] sentences that are identical apart from containing distinct coreferring names express the same communicated content; knowledge that my informant and I are participants in the same referential practice with a name does not in this case suffice to yield a sense in which we are same-sayers and same-thinkers.) Note too that such a descriptivism provides an immediate answer to the orthodox Millian’s objection that, if names are nothing more than tags (or “signs . . . of the mere objects,” as Mill [1889, 22] puts it), then anything other than what a name stands for, including anything descriptive, must be irrelevant from the point of view of semantics.\(^{11}\) The causal descriptivist’s response is that names can be associated with reference-fixing properties whose more or less idiosyncratic nature makes them entirely unsuitable as public meanings, but whose place in a conventional institution of naming makes it easily possible for others to latch on to associated idiosyncratic reference-fixing properties of their own, with reference to a common object the desired and likely result. In one sense, therefore, such properties are indeed irrelevant to the fundamental purpose of discourse involving names, which is to use a name to talk about a common object of reference that different speakers are differentially acquainted with. In another sense, however, the properties are pivotal: they, and not external causal-informational chains beyond the ken of speakers, are what determine reference to this common object.

Notes

1. See Kripke 1980. The classical forms of descriptivism attacked by Kripke are found in Frege 1952 and Russell 1905. Soames 2002 is a useful discussion of the different types of arguments (modal, epistemic, semantic) offered by Kripke, and of the main descriptivist responses to such arguments.

2. Kroon 2004b defends the reference-determining role of just such properties by using ideas found in Mill’s famous defense of the view that names have a denotation but no connotation.

3. See also Braddon-Mitchell 2004, which responds to the suggestion in Davies 2004 that it is best to see speakers’ capacities to recognize the referents of names as based on the grindings of a subpersonal system.
4. The aardvark case comes from Kripke 1980, 96, where, after citing the condition that the receiver of a name must intend to use it with the same reference as the person from whom he heard it, Kripke writes: “If I hear the name ‘Napoleon’ and decide it would be a nice name for my pet aardvark, I do not satisfy this condition” (ibid.). (An earlier discussion of the phenomenon of indirect reference borrowing occurs in Kroon 1987.)

5. Imagine, for example, that Smith’s mother named her son ‘Napoleon’ after Napoleon the pig in George Orwell’s *Animal Farm* (don’t ask why). If, knowing the mother’s penchant for recycling fictional names, Jones correctly believes that the name of both son and aardvark derives from the name of a fictional creature, he has beliefs that now give him the ability to use ‘Napoleon’ to refer to the fictional pig Napoleon as well (I put aside the question of how we are to understand fictional reference of this kind). Assuming he further believes that the author of the fictional work probably recycled the name of a famous historical person, he also has the ability to use the name to refer to Napoleon Bonaparte. In all these cases, the speaker manages to exploit certain causal-historical connections to his own referential ends. By contrast, Kripke claims in a footnote to his aardvark comment that “I can transmit the name to other people. For each of these people, as for me, there will be a certain sort of causal or historical connection between [their] use of the name and the emperor of the French, but not of the required type” (Kripke 1980, 96, fn. 43).

6. Soames’s objection is focused on a way of understanding such descriptions that is explicitly tied to Jackson’s and Chalmers’s two-dimensionalism (see, e.g., Jackson 1998a,b; Chalmers 2002a), which has as one of its “central characteristics” the view that “de re belief, knowledge, and the like, about individuals or kinds is to be explained as arising from the attitudes agents bear to indexical sentences, or mental representations, that express propositions [about] those individuals or kinds” (Soames 2005, 322).

7. See especially Soames 2005. It should be pointed out, however, that a number of causal descriptivists, including Jackson, share at least some of these doubts. They have on the whole been wary of uncritical talk of a meaning-equivalence between names and descriptions, in part because the reference-fixing properties that speakers associate with the same name can show such variation but also because it is unclear what causal descriptivists should say about the role of names in propositional attitude statements in particular. This makes it all the more surprising that Soames’s attack on Jackson’s views assumes that the latter takes proper names and natural kind terms to be synonymous with context-sensitive, rigidified descriptions.

8. But remember that even though classical descriptivism began with these ambitions, even Frege and Russell realized that the hype was a little overstated. The fact that it was far from unusual for names to have different descriptive contents for different language-users meant that communication often had to proceed in ways other than by speakers and audience grasping the same descriptive content, and it
meant too that the semantics of propositional attitude attributions had to be more flexible than their official theories made out (else we can't account for the truth of a sentence like ‘Everyone, including Scott himself, believed that he [Scott] was a marvelously inventive author’).

9. This part of the argument also occurs in Kroon 2004a, which responds to Jackson’s argument in Jackson 2004 that the reference-fixing properties associated with names contribute to what is communicated with sentences containing the names. The other Frege–Russell problems are likely to need rather different tricks for their solution. Chalmers 2002b, for example, uses the distinction between epistemic and subjunctive content to provide a novel descriptivist account of the way names behave in propositional attitude statements.

10. For an exposition and defense of doctrinaire Millianism, see, e.g., Salmon 1986 and Soames 2002.

11. Such a view has long been argued by Wettstein (see, e.g., Wettstein 1986, 193).

Are all the facts about nations, cultures, and economies really just facts about people’s mental states and their interactions? Are all of the properties that determine whether or not a thing is a work of art really just physical properties of that thing? Is linguistics, the scientific investigation of language, best understood as a branch of psychology, the scientific investigation of the mind? Can psychology be reduced to biology? Can all biological phenomena be explained chemically? Is chemistry really just part of physics? Is there anything going on in the world that isn’t a physical thing? Can there be freely chosen, autonomous human action in a purely physical world? Frank Jackson has made a controversial claim about the way in which one should investigate questions like these. This essay is a qualified defense of his claim; qualified because although it begins by defending Jackson’s claim against a range of objections, it concludes with a suggested modification to overcome an objection that we take to have force. Readers who are sanguine about Jackson’s chances of overcoming most of these objections might glance over our first six sections to see if they agree with us about these matters, but then should look at section seven more closely.

The questions are examples of nothing-over-and-aborvery: each of them asks whether the subject matter of one human classificatory practice or discipline is anything over and above the subject matter of some other, putatively more inclusive, more fundamental, or better understood human classificatory practice or discipline.

Lots of scientists and philosophers go in for nothing-over-and-aborvery. Because of where we are heading, it is worth noting that nothing-over-and-aborvery is a characteristic preoccupation of a certain breed of boffin,
the *serious metaphysician*, as Frank Jackson calls her (Jackson 1998a, 4–5). A serious metaphysical project is one that aims to produce, or at least explain why in principle it is possible to produce, a complete inventory of all the things or phenomena in some specified class by appealing to some restricted class of more primitive things or phenomena. For example, suppose your preferred ontological primitives are all physical. Then the job of the boffin who is interested in the mental and the social is to catalog and theorize about everything mental or social in purely physical terms, or at least to argue persuasively that this could be accomplished. The boffin’s challenge is to explain away the appearance that mental and social facts belong to a different ontological category: one that requires us to believe in minds and cultures *in addition* to atoms and forces. She is engaged in nothing-over-and-abovery. Jackson suggests that this challenge of locating one subject matter within another is what makes serious metaphysics both interesting and serious.

Nothing-over-and-abovery is one of the two central strands in our discussion. The other is *conceptual analysis*. Conceptual analysis is what we do when we reflect on whether rules can have exceptions, whether robots can be moral agents, or whether a fully scored chamber work for cellos and alto saxophone gets to count as jazz if it is written by the African-American composer Henry Thredgill. It is what German mathematicians did in the late nineteenth century when they thought about infinity and noted that two infinite collections of objects, unlike two finite ones, can be the same size even when one is only a subcollection of the other. We indulge in conceptual analysis when we reflect fairly systematically on some notion—like rules, moral agency, jazz, or infinity—which forms part of some human practice. Systematic reflection means, roughly, deciding which possible scenarios are scenarios in which we have something to which the notion applies, which possible scenarios are ones in which the notion is not applicable, and which scenarios are ones in which it is indeterminate. Thus, we distinguish cases of jazz from cases of nonjazz and both from borderline cases. As Grice put it: “If I philosophize about the notion of cause, or about perception or about knowledge and belief, I expect to find myself considering, among other things, in what sort of situations we should, in our ordinary talk, be willing to speak (or again be unwilling to speak) of something as causing something else to happen; or again of someone as seeing a tree; or again of someone as knowing rather than merely believing that something is the case” (Grice 1958 [in 1989], 172).

In line with this approach, Jackson explicitly takes conceptual analysis to be *the method of possible cases* (Jackson 1998a, 31–32).
Jackson has written a lot about philosophical methodology in the last decade. His efforts culminate, but do not terminate, in his 1998a *From Metaphysics to Ethics: A Defense of Conceptual Analysis*. He argues for the following thesis (for his formulation, see Jackson 1998a, chapter 2).

**Jackson’s thesis** Conceptual analysis is part and parcel of any nothing-over-and-abovey. In other words, the method of eliciting intuitions about possible cases is standard issue whenever we try to redescribe the subject matter associated with one kind of human classificatory practice in terms of the subject matter associated with another, more inclusive, classificatory practice. A fortiori, conceptual analysis is part of what needs to be done by the serious metaphysician: it is part of what you need to do to give a complete account of what sort of stuff there is in terms of an austere package of ontologically primitive descriptors.

This thesis has attracted a lot of attention. Some people have written about its role in arguments for particular philosophical claims. For example, David Chalmers, and in former times Jackson himself, have appealed Jacksonianly to intuitions when arguing that *qualia* cannot be explained physicalistically, so we often find criticisms of Jackson’s thesis tangled up with criticisms of nonphysicalism about the mental (see, e.g., Block and Stalnaker 1999; Byrne 1999; Yablo 2000). Commentators on Jackson have also seized upon the fine details of his approach to conceptual analysis and especially on his views about the nature of possibility and the a priori, which invoke the apparatus that Segerberg (1973) called *two-dimensional modal logic* (see the just-mentioned works, and also Stalnaker 2003, chapter 10).

We will argue that Jackson’s thesis is correct, but that his own formulations of it misdescribe conceptual analysis. Jackson, we think, wrongly assimilates conceptual analysis to the semantic analysis of ordinary language. We will keep the discussion very general and very methodological. We will not go into detailed case studies about the mind, the colors, ethics, or even jazz. We will also not engage with the details of Jackson’s two-dimensionalism. We will argue for Jackson’s thesis by presenting a series of objections coupled with replies on Jackson’s behalf. The objections are distillations of criticisms that can be found in the literature, but we use them primarily as dialectical tools for elaborating Jackson’s position and clearing up misunderstandings. Initially, our replies will be compilations or restatements of points that Jackson himself makes. Soon enough, though, our replies will exceed or depart from Jackson at least in emphasis. We cannot promise that Jackson would endorse our replies to the later
objections. We will end with the promised worry of our own—not about Jackson’s thesis as such, but about his preferred formulation of it. As noted above, we believe that Jackson misconstrues the nature of conceptual analysis. There will be no reply to this final objection; we think it succeeds.

Before we present the objections, a brief terminological digression for those, and only those, who are uncomfortable with our introduction of ‘nothing-over-and-abovery’. Traditionally, ‘reduction’ and ‘reducibility’ have been used where we use ‘nothing-over-and-abovery’. But which examples of nothing-over-and-abovery count as speculations about reducibility is controversial. Typically, reducing a theory \( T \) of the subject matter of some practice to some other theory \( T^* \) is taken to require that two criteria be met:

1. All the laws, equations, or other principles in \( T \) are derivable from or entailed by laws, equations, or principles in \( T^* \), perhaps in conjunction with intertheoretic bridge principles.
2. Every item in the subject matter of \( T \) is identical or nomically coextensive with an item, or is constituted by items, from the subject matter of \( T^* \).

However, there are theories \( T \) and \( T^* \) for which at least one of these criteria arguably fails and yet the contention that the subject matter of \( T \) is nothing-over-and-above the subject matter of \( T^* \) remains a live option. Consider, for example, the relationship between the mental and the physical. Davidson’s anomalous monism violates (1). Davidson (1980) says that there are no psychological laws and no laws linking psychological generalizations with laws of physics. Yet Davidson maintains that every mental event is a physical event, and this makes him a nothing-over-and-aboverist. Of course, it also means that he endorses (2), but his violation of (1) makes him a nonreductionist in the eyes of many. A variation on his view shows us how both the criteria for reduction can be violated without preventing the view from being an instance of nothing-over-and-abovery. Because it is an anomalist view, it violates (1). However, this variation allows that there is no single physical state, event, or cluster of states or events that is identical or coextensive with or that constitutes my belief that strong leadership is wrong leadership. To say that I have this belief is to report on certain consequences of facts about my total mental state, namely actual or possible interactions between my brain, certain input stimuli, and certain output behaviors. Hence this anomalist holist view also violates (2). (To make the violation even more extreme, make the holism highly externalist:
whether my total mental-state/mental-processing is such that I believe that \textit{strong leadership is wrong leadership} depends both on features of my present environment and on the details of how my brain came to be in its present configuration.) All of this is compatible with the claim that the mental is nothing over and above the physical. Our holist (externalist) anomalous monist is a nothing-over-and-aboverist.\footnote{Because the requirements on reduction are typically taken to be more stringent than those that concern us in what follows, we will stick to our own term, ‘nothing-over-and-abovery’. If you dislike it, you are welcome to substitute ‘ontic dependence’, ‘ontic determination’, or even ‘reducibility’. If you have a developed account according to which the idea can be captured in terms of supervenience theses, then make yourself at home. }

2 First Objection: The Record of Forays into Conceptual Analysis Is Dismal

It is well known that attempts to capture the rules we implicitly follow when we attribute \textit{knowledge} to somebody, \textit{meaning} to an utterer, or \textit{causation} to a sequence of events have run out of steam because they have failed to produce satisfactory analyses. Famously, Edmund Gettier (1963) challenged the conceptual analysis of \textit{knowledge} as justified true belief. He exhibited counterexamples to it, and these prompted new definitions, which generated new counterexamples. Nobody has yet proposed an analysis that satisfies the philosophical community. Similarly, Paul Grice (1957) offered an account of what it is for an agent \textit{S} to mean that \textit{p} on a particular occasion when uttering \textit{x}. This reconstruction of Grice’s original definition comes from Schiffer:

\begin{quote}
\textit{S} means that \textit{p} in uttering \textit{x} iff, for some person \textit{A} and feature \textit{Φ}, \textit{S} intends

\begin{enumerate}
\item \textit{A} to recognize that \textit{x} has \textit{Φ};
\item \textit{A} to think, at least partly on the basis of thinking that \textit{x} has \textit{Φ}, that \textit{S} uttered \textit{x} intending \textit{A} to think that \textit{p};
\item \textit{A} to think, at least partly on the basis of thinking that \textit{S} uttered \textit{x} intending \textit{A} to think that \textit{p}, that \textit{p}. (Schiffer 1987, 243)
\end{enumerate}
\end{quote}

Like the justified true belief analysis of \textit{knowledge}, Grice’s already convoluted definition of speaker meaning (as it came to be called) fell prey to sundry counterexamples, many of them Grice’s own, and spawned increasingly long-winded definitions until well into the 1980s. Likewise, David
Lewis’s counterfactual analysis of causation (see Lewis 1986b, chapter 21, “Causation”) generated increasingly complicated definitions with more and more epicycles (see Menzies 2001).

The definitions that typified these projects appeared more and more unnatural as time passed and the counterexamples kept on a-comin’. We can argue by induction that there is no point in pursuing these projects any further, and we can speculate that something about the way our minds operate or about the nature of the concepts means that definitions just are not available for them. The prospects for successful analysis seem equally grim in other areas of philosophy, so we should conclude (albeit tentatively) that this lesson extends to analyses of belief, desire, the Good and the rest. We should give up on conceptual analysis of the kind that has filled the journals (especially Analysis) with definitions and counterexamples. Instead, we should channel our energies into working out what it is about the mind or the concepts that gives them this elusive character. If all of this is right, then Jackson is wrong to hold out any hope that conceptual analysis can pay dividends as a tool for assimilating one subject matter to another, or as a tool for anything at all.

2.1 Reply

Note first that much successful conceptual analysis occurs quietly out of the limelight. If I tell you that you are doing all right because your home is worth $700,000, you might well complain. On analysis, the claim that your house is worth $700,000 means, roughly, that if you sold it now you could expect to get about $700,000. But you don’t want to sell it. Your rates have gone up. You’re not doing all right at all.

Some successful conceptual analysis is not so workaday. An example of this is our earlier one from mathematics of successful (albeit partial) efforts to tame the concept of the infinite within a tractable mathematical theory.

True, the projects that have flooded the pages of the philosophy journals have not been completed successfully. But they do concern philosophically intractable notions and we shouldn’t expect the job to be easy. It is precisely because these projects are long and hard that they take up so much space in the journals.

Complete analyses of even these hard philosophical chestnuts must be available in principle. After all, if you describe a possible scenario to your informant in enough detail, she can tell you whether the concept as she uses it applies or not or whether she is unsure. Tell her a story about something painted by a robot and hung in an art museum and she will have
something to say, however noncommittal, about whether or not it’s art—
even if all she says is that she’s not sure. Tell her a story about a planet
where the clear, drinkable liquid that fills the oceans is not H2O but some-
thing else, and she will have something to say about whether or not this
liquid is water or why we can’t decide. Indeed, the reason we know that
most of the post-Gettier definitions of knowledge have counterexamples
is because our informants can identify cases where the definition is satisfied
but where there is not knowledge, or vice versa. All of this means that
there is in principle a story to be told about our informant’s responses to
all possible cases, and this story is the right analysis of the concept. It will
say, for every one of infinitely many possible scenarios, the degree to which
one takes it that the concept applies in that scenario. Not an elegant defi-
nition, maybe, but an analysis of the concept as our informant employs it
nonetheless.

Although any complete analysis of a concept has to be a statement of
necessary and sufficient conditions for its application—a definition, if you
like—it might not be a definition of the sort one would naturally wish for.
For a start, we should not be surprised if some of the rules we implicitly
follow when classifying things as jazz or nonjazz, right or wrong, art or
nonart, water or nonwater, are horrendously complicated and gerryman-
dered. It would be marvelous if the quest for a conceptual analysis often
terminated in a more or less pithy statement of necessary and sufficient
conditions like Grice’s candidate for a definition of awe: “a mixture of fear
and admiration” (Grice 1958 [in 1989], 176). Unfortunately, however, our
informal, everyday classificatory practices evolve over time in response to
changing circumstances and are not subject to careful review by the phi-
losophy department or the Crown Law Office. The question of whether to
include a particular musical event among the jazz may well depend on
vagaries: what were the journalists working for *Downbeat* magazine most
interested in during the late ’50s? What kinds of music did most of the
jazz players from the ’30s move on to when the big bands disbanded?
Perhaps the best definition of the concept *jazz* will be a motley disjunction
of conditions. And perhaps the best definition of the concept of *meaning*
will be of the long-winded Gricean type illustrated above with exceptions,
caveats, and epicycles onboard.

We should also not be surprised if a complete analysis is noncommittal
on certain matters. There are plenty of ordinary users of ‘water’ who genu-
inely have no view about whether XYZ counts as water on Twin Earth,
and maybe a correct analysis of *water* (as the concept is used by those
speakers or by the whole community) should reflect this.
These points about what we can expect from a definition suffice to undermine the usual antianalysis complaint that increasingly complex definitions spell increasing likelihood that the search for a complete account is faltering. Hedged, complicated analyses might be just what we should expect.

Nevertheless, an objector to this reply might rejoin, there’s something wrong somewhere with the yarn we have been spinning. There must, she might insist, be some concepts for which complete analyses are in principle unavailable. Otherwise, an unthinkable infinite regress of definitions looms—unthinkable because of natural language’s expressive limitations.2

To meet this worry, we concede that there might well be concepts that resist complete analysis in languages with finite expressive resources, and our spade may indeed be turned by some of these if we keep defining complex concepts in terms of simpler concepts for long enough. But consider one more point about completed analyses. We should not be too surprised if some completed conceptual analyses resist formulation in finite languages. Suppose we are analyzing mental states. We ask: “what is it for something to count as the belief that strong leadership is wrong leadership?” We ask, in other words, which of the infinitely many possible states of the world are those about which an ordinary, folksy wielder of the belief concept would claim that I hold this belief. If we take seriously the idea that conceptual analysis is the method of reflecting on (the infinitely many) possible cases, we may discover that only an infinitely long list of the possibilities to which the concept applies captures the rule governing the concept’s use. We might hope for better than this, but it may not always be obvious that optimism is warranted. This recourse to infinitely long lists of conditions will surely apply too in the case of any concepts whose unpacking would take us beyond the expressive resources of finite languages.

Since completed analyses need not yield satisfying, quotable sound bites, conceptual analysis must not be identified with what Stephen Stich calls the “method of proposing definitions and hunting for intuitive counterexamples” (Stich 1993, 354). To be sure, a fun way to elicit intuitive judgments about whether a concept (jazz, for instance) applies to a possible case or not is to entertain a definition of it and search for counterexamples. However, conceptual analysis, as Jackson understands it, is the process of reflecting on possible cases, and this need not, though it may, consist in auditioning, testing, refining, rejecting, and replacing definitions.

Our original objector has one more worry. If complete analyses of many philosophically interesting concepts are so elusive, what’s the point of the enterprise? Why go in for analysis if there’s often little hope of finitary completion?
Our reply: We don’t always need a complete analysis of a concept in order to learn something worthwhile about it from the business of analyzing it. Failure to net the necessary and sufficient conditions for the application of some concept to a circumstance need not spell a failure to learn more about the application of the concept. It might still be of interest to uncover, as we have done in the case of knowledge, necessary but insufficient, or sufficient but unnecessary, conditions for its application, and we should not regard the unavailability of even provisional definitions of some concept as a sign that we are getting nowhere with the analysis of it. Philosophers probably know much more about knowledge as a result of the discarded analyses thrown up in Gettier’s wake. Much the same can be said about our insights into meaning, causation, and especially infinity, which owe so much to incomplete or downright incorrect analyses.

In sum, conceptual analysis, regarded as the probing of our intuitions about the application of some notion in various possible scenarios, is alive and well: it makes unobtrusive contributions to scientific, philosophical, and ordinary reasoning. It pays dividends even when it does not converge on an exceptionless criterion of application for a concept. So for all our objector says, it is available to play its allotted role in Jackson’s program.

3 Second Objection: Armchair Reflection Is Irrelevant to Finding Out What the World Is Like

There is something odd about the idea that one can sit, secluded from life’s laboratory, and make substantial progress toward discovering what the world is like or what sort of stuff it contains. This is what we learned when we finally realized that natural science relies primarily on a posteriori discoveries. To flesh this out a bit (and this is how Jackson himself puts the objection in Jackson 1994b, 94) the appeal to intuitions about possible cases is doubly irrelevant to finding out what the world is like. First, it is an appeal to intuitions, not to facts. Second, it is an appeal to the possible, not the actual. We could augment this objection by recalling, third, that Kripke (1980) and Putnam (1975) made a big deal of the fact that much that is necessarily true—such as the fact that gold is an element whereas salt is a compound—is learnable only a posteriori. In general, access to the nature of things is not a priori access.

3.1 Reply

Part of what it takes to discover whether there are Ks and what they are like is determining what it would take for us to regard something as a K. This latter is a job for the conceptual analyst. To explain: A serious metaphysician,
or any scientist, if it comes to that, wants to know whether the world contains certain things—mesons, gods, minds, values, world souls, or whatever. She also wants to know what these things are like. Hers is the project of discovering whether there are any Ks, which things (if any) are the Ks, and what (if anything) the Ks are like. However, the methodological starting-point for any such project is the question: Under which circumstances would we judge that there are Ks? Only by shedding light on this question can we proceed, since only then can we go out into the world and discover whether any of the circumstances in which we would judge that there are Ks obtain. And only then can we say whether there are Ks, which things are the Ks and what the Ks are like. This does not mean that we must completely settle the question of what we would regard as a K before we write the surveys and perform the experiments. The two complementary projects—a priori conceptual analysis and a posteriori empirical examination—might interlock and cooccur. If conceptual analysis has any priority over empirical investigation, it is logical or methodological, not temporal.

Our objector will already be restless. It is not obvious that we need to know what we would regard as a K, as a god, for example, in order to look for one. Maybe we need to know what we would regard as God in order to go out and look for the sort of thing that we would regard as God. But a search for God is different from a search for what we regard as God. We are interested in whether there are real gods, real mesons, and so on. This is not the same as being interested in whether there are items that we would regard as gods or as mesons.

Here’s the response. We can’t embark on a search for Ks unless we have some idea what we are looking for. As Jackson says:

When bounty hunters go searching, they are searching for a person and not a handbill. But they will not get very far if they fail to attend to the representational properties of the handbill on the wanted person. These properties give them their target, or, if you like, define the subject of their search. Likewise, metaphysicians will not get very far with questions like: Are there Ks? Are Ks nothing over and above Js? And, Is the K way the world is fully determined by the J way the world is? in the absence of some conception of what counts as a K, and what counts as a J. (Jackson 1998a, 30–31)

Sometimes what we are looking for will be obvious, and we hardly need to glance at the handbill. If it is marbles, then we already know what marbles are, so we can just go out there and look for them. If it is mesons, well, we might have to consult our scientific theory, but often we will have done that long before we started looking for the mesons, so it’s as good as done. However, if it’s God we seek, we might have to consider carefully
the sort of thing we are looking for, methodologically prior to the search. The only reason that it makes any sense to go out and look for a God is that we have a practice of talking and thinking about God and of fine-tuning our behavior depending on whether it seems to us that there is a God or not. The search for God is, at least initially, a search for the thing that meets our God-profile. This is why the search for God confronts us early on with the question: “What would a thing have to be like before we would regard it as God?” To get any sort of fix on our target, we must have a conception of *Godhood*—of the circumstances under which we would ascribe it—that informs our search, even if the conception is something we develop in the process of the search.

Now we can see how conceptual analysis comes in at the methodological starting point. This question “What would a thing have to be like for us to regard it as God?” just is the question “Under what clearly specified circumstances would we say that we have a God, which situations would clearly be Godless, and when don’t we have a clue?” And that is just a question about possibilities—one we can ask and answer from the armchair so long as we have a rich enough description of the cases on which we are to rule. (We might ask, for instance, whether a circumstance in which the world was created by a mortal but omniscient mouse would be one in which God created the world. Our answer would determine whether, should we discover that this was indeed the case, we would say that God in fact created the world.)

What has all this to do with nothing-over-and-abovery? Why is conceptual analysis needed there? Well, questions of the form ‘What does it take for something to be regarded by us as a K?’ always crop up when we are considering whether and to what extent the subject matter of one classificatory practice can be assimilated to that of another. Suppose you hold the view that economic facts are just constructs made out of psychological facts: you believe that the economic is nothing over and above the psychological. If you are ‘seriously’ going to claim that posits like supply and demand are constructs of the psychological, you had better have to hand a story about how to translate between theories within psychology and the economic theories in which supply and demand feature. And to do that you need to know something fairly detailed about the role played in economic theory by the posits. You need to know what it would take for us to regard something as a fall in demand or as a rise in supply. In particular, you need to know which scenarios, described in austere psychological terms, count as falls in demand, which as rises in supply, and so on.
Or suppose you are wondering whether freely chosen action is possible in a world where every action is causally determined by the physical facts up to the moment that the action is performed. You are wondering whether, in such a world, free choice is anything over and above something physical. To approach the matter, you must consult intuitions about which acts are free and which are not. In particular, you need to consider whether there are any physicalistically determined acts that would count, according to those intuitions, as free acts.

Now we can answer the original objection point by point. First, yes, conceptual analysis appeals to intuitions, not to facts, but intuitions about cases are highly relevant to the search for empirical facts. They are relevant because it is only when we have an intuitive idea of what sort of thing we seek that we can go out and seek it. As Grice has it: “It is a very old idea in philosophy that you cannot ask, in a philosophical way, what something is unless (in a sense) you already know what it is” (Grice 1958 [in 1989], 173). Second, yes, conceptual analysis appeals to the possible, not the actual, but appeals to merely possible cases are relevant because we conduct our search in partial ignorance of the way things actually are; the point of the search is to see whether one of the possibilities in which there are Ks is the actual case. Third, yes, access to the nature of things is not in general a priori access, but the role we have specified for conceptual analysis need not have anything to do with the discovery of essential natures. The discovery that gold is by nature an element was indeed an empirical discovery. Jackson’s point, however, is that we were only able to make it because we knew, by consulting our intuitions, what sort of actual shiny, valuable thing with samples of which we are acquainted would be regarded by us as gold. The intuitive knowledge that gold is shiny and valuable enables us to pick up the scent, and we need some way of picking up the scent; so conceptual analysis is needed for the research. Nobody need claim that gold’s shininess and comparatively high value are essential properties of it. They are just properties we need to know about in order to strike gold.

4 Third Objection: Jackson’s Description of Scientific and Philosophical Discovery Is Inaccurate

A common worry about Jackson’s story is that assimilations of one scientific subject matter to another do not in practice require sustained or systematic reflection about possible cases in the light of our intuitions. Instead, they are a combination of a posteriori discovery and ontological
parsimony. Consider, for example, the discovery that the temperature of a gas is simply the mean kinetic energy of the molecules that constitute it. The story is sometimes told this way. The kinetic theory of gases was a theory about what gases were made of: it said they were made of small molecules. Within this theory, it was possible to explain the fact that a gas had a certain temperature in terms of the motion of the molecules. This was an empirical discovery. But then general methodological considerations about how science should be done kicked in. We want our overall picture to be uncluttered. It is far simpler to rule that the mean molecular kinetic energy of a gas is its temperature than to opt for a more complicated theory in which we distinguish mean molecular kinetic energy from temperature.

4.1 Reply
Ontological extravagance was never an option. Here’s an alternative account of this example which invokes Jackson’s thesis. The temperature of a gas is a posit in the thermodynamic theory of gases, but that theory does not pretend to be a fundamental theory. We can consider various possible cases in which gases are constituted in various ways and ask, for each such case, what would count as the temperature of a gas, if anything. Of course, we don’t need to run through all these cases in advance of doing some more fundamental science, but when we discovered that gases are composed of independently moving particles, we knew enough about the commitments of the less fundamental thermodynamic theory to know that mean molecular kinetic energy does the work that temperature does in that latter theory.

The advantage of this Jacksonian account is that it doesn’t require an appeal to simplicity. Once we knew the molecular dynamics of gases, we knew what temperature really was. It was not like deciding whether we really needed ungainly mechanisms to make our theory work. It was just a case of redescribing a macrophenomenon in microphenomenal terms.3

5 Fourth Objection: Metaphysics Should Not Be Hostage to Conceptual Analysis

Surely it’s okay if, in the end, the things in the world that turn out to be the Ks lack many, perhaps even most, of the characteristics that our unsophisticated folk theories attributed to Ks. Bill Lycan is interested in showing that mental states like beliefs and desires are really just physical. He is a nothing-over-and-aboverist about the mental, but he doesn’t like the Jackson thesis: “I am entirely willing to give up fairly large chunks of our
commonsensical or platitudinous theory of belief or of desire... and
decide we were just plain wrong about a lot of things, without drawing
the inference that we are no longer talking about belief or desire” (Lycan
1988, 31). If Lycan is right then, in the case of beliefs and desires anyway,
it’s not such a big deal which kinds of physical scenario would count for
us as cases where we have a belief or a desire. Lycan suggests that it doesn’t
really matter if we don’t pin down our subject matter by analyzing our
practices of ascribing mental states.

5.1 Reply
Lycan can’t really be right. If what you discover is fairly unlike what you
started looking for, then either you have changed the subject or else your
concept evolved while the search was underway (and perhaps as a result
of the search). If it really is belief or desire that we are interested in assimilating
to the physical, it seems to matter a great deal that we not stray far
from what the people who talk about beliefs and desires have in mind.
Otherwise, it seems that we are no longer talking about beliefs and desires
(see Jackson 1998a, 37–41). Of course, what the folk regard as criterial for
something’s being a belief or a desire is almost certainly a moveable feast.
Different folk will have different views about it; most folk are unreflective
about it; and the impact of scientific discovery, social evolution, apparent
inconsistencies among the criteria, or just new kinds of experience may
prompt radical criterial revision. We cannot assume that our concepts are
uncontested, determinate, or stable over time. And of course, if we haven’t
done the conceptual analysis yet, we won’t know just which things the
folk would be prepared to count as beliefs or desires. We also won’t know
what changes they would be prepared to make to their concepts of belief
and desire in response to scientific investigation.

In this connection it is interesting to note what Lycan goes on to say:
“I think that the ordinary word ‘belief’ (qua theoretical term of folk psy-
chology) points dimly towards a natural kind that we have not fully gra-
sped and that only mature psychology will reveal” (Lycan 1988, 32). This
might well be right. But if it is right, it is a piece of accurate conceptual
analysis. It is right because although the folk are fairly elastic about what
they will count as a belief when a situation is described to them in physical
terms, they will probably want beliefs to be fairly natural physical kinds.
They will probably weight the intuition that a belief is a natural kind very
highly and be prepared to say that they were wrong in lots of ways about
what it meant to believe something so long as beliefs are fairly natural
things, according to the best physical ontology. That’s if Lycan’s right. In
other words, Lycan is wrong to think that it’s okay for the theorist to give up on large, weighty chunks of folk theory about beliefs and still say we have beliefs, but it might turn out that the very things he wants to give up on are such that the folk will let him be their guest.

Having said all of this, there’s surely something right in what Lycan says. Surely science need not be subject to the tyranny of our a priori conceptions of things. Here’s where we start making claims Jackson might not endorse while still, we think, defending his thesis that conceptual analysis is needed for nothing-over-and-abovery. Consider another objection.

6 Fifth Objection: Philosophy Rightly Understood Is Continuous with the Natural Sciences, and Therefore Does Not Require Conceptual Analysis

Kim Sterelny once wrote: “My approach [to the mind] is not just physicalist, it is naturalist. Naturalists are physicalists. . . . [B]ut naturalists have methodological views about philosophy as well; we think philosophy is continuous with the natural sciences. On this view, philosophical theories are conjectures whose fate is ultimately determined by empirical investigation. . . . An alternative conception is to see philosophy as an investigation into conceptual truths proceeding by thought experiments that probe the way we understand our own concepts” (Sterelny 1990, ix).

6.1 Reply

First, Sterelny’s characterization of the second, less-than-naturalistic alternative philosophical program is not a useful characterization of Jackson-style nothing-over-and-abovery. Jackson does think we must investigate conceptual truths in order to nothing-over-and-aboverize, but, as we have seen, the role he assigns them is only that of providing a methodological starting point for our empirical research. Second, it’s not clear that there is any useful question to address about which work counts as philosophy and which counts as natural science. Yes, Jackson wants to say that philosophers should do conceptual analysis, but he also thinks that scientists should (and do) too, and he says nothing about whether there is anything else for a philosopher to do.

These preliminaries aside, though, we agree with Sterelny, at least to this extent. Surely you can put forward an empirical hypothesis about the sorts of phenomena dealt with in a high-level theory without being a hostage to that theory. There are projects that are naturalistic in Sterelny’s sense. However, they are different from nothing-over-and-abovery, so Jackson’s
thesis does not apply to them. And not all major debates in philosophy and science are instances. (Sterelny 2003 offers a revised assessment of the methodological situation which is in line with our remarks here.)

To illustrate the difference between Jackson-style nothing-over-and-above or on the one hand and Sterelny-style naturalism on the other, suppose I am interested in the mind: in what we are going on about when we talk of beliefs and desires, what consciousness is, and so on. Suppose I think too that there is nothing in the world but physical stuff and I want to account for all this physically. Well, one option is to do some nothing-over-and-above: to work out what, if anything, the ordinary people are doing when they ascribe and entertain beliefs and desires or manage to stay conscious. This is surely a worthwhile project: mental states are discussed, considered, and apparently experienced by ordinary people who also invoke them to explain and predict others’ behavior. If you buy into this project, we have argued, you need to do conceptual analysis. But there’s another way to go. You start by thinking about some interesting puzzle that arises within a physicalistic framework for humans and for other animals. How do these creatures manage to perform in a way that shows they are tracking the environment, planning ahead, responding to certain stimuli and ignoring others, grouping perceptual information together in particular ways, and so on? You could begin to construct a physicalist theory of how all this is done. You know, of course, that people attribute mental states to each other and to animals as well, and you might use the data about when they do this, and what they attribute, in the construction of your physicalistic (your naturalistic) theory. But your job is not to discover the extent to which these folk-psychological practices and their mentalistic ontology are grounded in the physical. You are not trying to redescribe or rehabilitate folk psychology. You are trying to solve some puzzles physically, and these happen to coincide more or less with the sorts of puzzles which we ordinarily explain in terms of our theory of the mental. You would therefore do well not to say that you have developed an empirical theory about the nature of belief—a theory that might tell us what belief truly is. Before you are entitled to do that, you must amalgamate your project with one that involves analysis of the relevant folk concepts. As Sterelny (2003: ix) puts it, you would need “both a well-developed account of . . . folk commitments about belief, and a theory of reference for folk psychological vocabulary telling us the extent to which folk psychology’s vocabulary depends on the accuracy of folk psychology’s picture of the mind.”

As Sterelny says, any theories you develop in the course of this investigation are empirical conjectures. This means, of course, that they are subject
to the assessment procedures that apply to any empirical claims in the natural sciences. If Lakatos (see Lakatos and Zahar 1978) is right, for instance, these empirical conjectures had better produce novel predictions soon, and those novel predictions had better be confirmed, or the scientific community will want to chuck your theory on the scrap-heap even if you don’t.

We think that projects of this naturalistic kind are fairly common. Take mainstream Chomskian linguistics, where a grammar is an empirical conjecture about the native speaker’s linguistic competence: a theory of the internal structures that correspond to the rules she has mastered. Such a theory is primarily a hypothesis about which onboard rules enable the speaker to understand or produce the infinitely many sentences she can understand or produce, and it is part of a psychological explanation of how those rules could have been mastered. It is not an attempt to collate what a lay native speaker regards (or could be made to regard via intuition probing) as the rules of her language. Even so, gathering the raw data for such a grammar involves eliciting folk responses to questions about possible expression-types: “Is this one grammatical?,” “Are these two synonymous?,” and so on. The aim in pumping these intuitions is not to produce an analysis of the folk concepts of grammaticality, synonymy, and the like. Still less is it to show that such concepts are nothing over and above psychological processes. Rather, it is to amass fallible information for use in the construction or testing of a theory about what rules the native speaker has internalized. Although such a theory is called “a grammar,” mainstream linguists do not take it for granted that the rules of their theory must be rules that the native speaker could be made to endorse. Intuitions are merely empirical evidence that certain rules are among those that have been mastered. We do not claim that grammatical theories developed within a broadly Chomskian framework rely less for their plausibility on folk intuitions than, say, theories about the nature of folk psychology do. Surely they don’t. Folk psychology, or folk ethics, or the folk account of free will can be theorized about empirically just as the internalized rules that explain our linguistic competence can. Our claim is rather that Chomsky’s starting point, the puzzle of linguistic competence, identifies his project as distinct from that of a nothing-over-and-aboverist. Chomsky explores folk intuitions in order to explain a folk capacity physicalistically. The physicalistic nothing-over-and-aboverist explores folk intuitions in order to investigate the physical natures of things the folk talk and think about.

Other examples of Sterelny-style naturalism are Ruth Millikan’s analysis of mental content in terms of proper function (Millikan 1986) and Michael
Devitt’s semantic program (Devitt 1996). They are plausibly not attempts to reduce the folk theory of mental content attributions or of meaning to something more ontologically pristine. However much conceptual work is involved in their construction, Jackson’s thesis is not in play.

7 Sixth Objection: Jackson Should Not Equate Conceptual Analysis with Investigations into Semantic Competence

To give the background of this last objection—the one we endorse—it will be useful to recapitulate, in Jackson’s words, what his thesis commits him to:

Serious metaphysics requires us to address when matters described in one vocabulary are made true by matters described in another. But how could we possibly address this question in the absence of a consideration of when it is right to describe matters in the terms of the various vocabularies? And to do that is to reflect on which possible cases fall under which descriptions. And that in turn is to do conceptual analysis. Only that way do we define our subject—or, rather, only that way do we define our subject as the subject we folk suppose is up for discussion. (Jackson 1998a, 41–42)

As noted earlier, our objection is not to Jackson’s thesis as such. It is that when he explains it his own way, he smuggles language into his characterization of conceptual analysis. We have talked about the subject matters associated with different human classificatory practices. Jackson talks about “stories told in different vocabularies.” He adopts an unashamedly linguistic outlook on conceptual analysis, and he is in illustrious company. In the passage from Grice quoted earlier—the one where he championed conceptual analysis—he talked about our talk: our talk about cause, about perception, about knowledge. Linguistic analysis has been central to philosophizing in many traditions throughout the twentieth century. Even so, Jackson would be better off without the linguistic turn, or so we will argue. Let’s look at more of what he says and then at why it creates unnecessary problems for him.

Our subject is really the elucidation of the possible situations covered by the words we use to ask our questions—concerning free action, knowledge, and the relation between the physical and the psychological, or whatever. I use the word ‘concept’ partly in deference to the traditional terminology which talks of conceptual analysis, and partly to emphasize that though our subject is the elucidation of the various situations covered by bits of language according to one or another language user, or by the folk in general, it is divorced from considerations local to any particular
language. When we ask English users in English for their intuitive responses to whether certain cases are or are not cases of knowledge, we get information (fallible information . . .) about the cases they do and do not count as covered by the English word ‘knowledge’. But our focus is on getting clear about the cases covered rather than on what does the covering, the word, per se. We mark this by talking of conceptual analysis rather than word or sentence analysis. (Jackson 1998a, 33–34)

We extract three claims from this passage.

(1) When Jackson talks about conceptual analysis, he is not really talking about the analysis of concepts. Instead, he is talking about how to understand the way that linguistic expressions function.

(2) The sense in which he is interested in how linguistic expressions function can be explained along the following lines: Those of us who use these words to pick out things in the world—things with particular properties—are able to do so because of certain beliefs, knowledge, or dispositions, which count as determining a rule for the use of the term. Maybe we can’t articulate that rule, but there must be one, and it must be one we have internalized somehow because we are able to say—given a full description of a possible case—whether the term would apply to that case or not. So when we say we are doing conceptual analysis, what we are really doing is looking for the knowledge, disposition, or whatever that determines, or constitutes, the rules governing our uses of terms—of the sorts of terms we find in metaphysics and science especially. We can think of this as the project of finding out what makes us semantically competent with these terms.4

(3) That isn’t quite the right characterization of conceptual analysis. The English word ‘water’ and the French word ‘eau’ and all the other translations of these words into various languages pick out the same kind of stuff. Strictly, it’s not the word ‘water’ (or ‘cause’, or ‘person’, or ‘belief’, or . . .) whose use we are investigating. Rather, it is something that each of these words has in common with its translations into other languages. This does not really alter any of what was said in (1) and (2), because if two terms really are exact translations of each other, the same rule is required for semantic competence with either—each is usable as a description of the same contexts as the other.

We have two worries about all this. First, Jackson’s idea about the nature of conceptual analysis is hostage to his own theory of semantic competence with words, his theory that I have access by mere reflection to the knowledge in virtue of which I understand a verbal usage. Chomsky denies that there is such a thing as semantic competence, and there are other
broadly Chomskian semanticists, like Richard Larson and Gabriel Segal (1995), who think that we do have knowledge that makes us semantically competent, but unlike Jackson, they believe it is locked away in a semantic module. For Larsen and Segal, this is the knowledge of how to prove the T-theorems of a truth theory, and it is not accessible to any kind of reflection. Jackson is welcome to his theory of semantic competence, but it seems as though conceptual analysis ought to make sense independently of who is right about semantic competence with words.

One gesture toward fixing this problem, while sticking to Jackson’s linguistic take on conceptual analysis, involves remembering that it is how we would use words in various scenarios that matters to conceptual analysis. Would I call that object with the lovely eyes and the beautiful singing voice a person, even if I knew it was a computer, or a Martian? The answer might depend a lot on who I am talking to, what information we share, whether I need or ought to be precise, polite, and so on. I won’t refrain from calling the computer a person if my refraining will confuse or horrify you. Specifying all this stuff is part of specifying the sort of case we are ruling on. This sounds as though it should be dealt with as part of pragmatics—of how we do things with words—rather than within the theory of what enables us to master and understand the meanings of words.

Our second worry is this. When we are interested in what it takes for an action to be right, or what causes what, or who are the people, it really doesn’t seem that we are interested primarily in what rules we follow in virtue of which we understand the words ‘right’, ‘cause’, and ‘person’. For one thing, it may be that a being without language can distinguish people from nonpeople, can act in a way that requires an understanding of cause and effect, and so on. Such a being would be operating with the same concepts as the ones we have words for. For a second thing, there might well be philosophically interesting concepts for which we lack words. For a third thing, even among the linguistically endowed, verbal conduct is only one of the behaviors that manifest our conceptual commitments. Who are the persons is evidenced not only by when we are disposed to call them persons but also by the fact that we wonder whether they should marry our daughters. What counts as a vote for Bush is evidenced not only by what gets called a vote but also by which chads we don’t leave hanging. Indeed, if we wanted to find the translation into Jungle of the English word ‘person’ we would look to see what gets punished, what gets married, and so on—not merely because these are useful nonverbal signs of which word we are looking for but because the word, whatever it is, normally gets used within a comprehensive person-tracking practice. Now perhaps some of
these tell-tale nonverbal activities and the kinds of thinking they require would not be available to a nonlinguistic being; perhaps the acquisition of many concepts requires linguistic competence. Still, it doesn’t follow even from this that what we are interested in analyzing are the words we use. Finally, suppose that despite the above misgivings, conceptual analysts steadfastly restrict their attention to linguistic data. Still, one need not employ or even possess a word that covers all and only the persons in order to betray linguistically that one has a concept of persons. Remarks like ‘Nobody alive can remember the Crimean War’, ‘If you object to abortion, you should object to euthanasia’, and ‘War crimes should be punished’ speak volumes about person concepts, whether or not our language has a word that translates ‘person’.

All in all, we should not regard the objects of analysis as linguistic terms. Analyzing our talk may be the best way to analyze our classificatory practices, but surely it is really only a heuristic. The practices themselves—or the concepts or dispositions that generate them—are our real quarry. True, we have not really said what it is we are analyzing when we do conceptual analysis. We have said only that it is not language. But successful conceptual analysis can and does proceed without a clear fix on what concepts are, just as successful arithmetic proceeds without a clear fix on what numbers are. We should not assume that the success of conceptual analysis and the fact that we owe much of this success to our facility with the meanings of words deliver the verdict that words are our analysandum. This verdict is corroborated by Jackson’s admission that it is the semantic rules shared by different languages, not the words themselves, that really interest conceptual analysts. If we are interested in the common background machinery that is shared across languages, then, yes, we might invoke very similar semantic rules in order to be competent with terms in different languages, but surely this is only because these different words are all getting at the same underlying thing—the common practice or the common concept.

We end by noting two likely reasons for Jackson’s assimilation of conceptual analysis to an investigation into natural language semantics. The first reason is that linguistic tokens are admirably public, inspectable, and physical. Even so, they are not what we are really after, and we should not pretend that they are just because the job of finding what we really are after is hard. The second reason is more interesting. One thing that Jackson’s linguistic focus enables him to do is talk about our classificatory practices as though they were collections of sentences—roughly, as though they were theories in the logician’s sense. He talks about ‘belief’ as a kind
term in a theory called ‘folk psychology’, just as one might say that ‘temperature-in-a-gas’ is a kind term in the thermodynamic theory of gases. When Jackson describes conceptual analysis in a piece of nothing-over-and-above—let’s say in a quest to assimilate the mental to the physical—he likes to think of it as a process of stating, for every case described in the language of a physical theory, what (if anything) would be the right thing to say in the language of folk psychology. This enables him to think of the method of cases as ideally involving a translation between the language of one theory and the language of another, more pristine theory. For example, there is a long true statement in the language of physical theory which entails the statement that Fred believes that strong leadership is wrong leadership. This way of thinking of our practices as theories and of our theories as linguistic objects is very useful to Jackson. It immediately licenses him to discuss questions about entailment and consequence that arise when we consider the relationships between different theories—say, physics and psychology—and it lets him use as much formal apparatus for alluding to the inferential relationships among sentences as he likes. He can (and does) model the reduction of one practice to another by appealing to Ramsey sentences (see, e.g., Jackson 1998a, chapter 6, where he draws on work in Lewis 1970). He can elaborate his conception of possibility by invoking the two-dimensional modal formalism developed by Davies and Humberstone (1982) (see Jackson 1998a, chapters 2 and 3). These formal procedures have been rigorously presented over the years and have become well understood, in just the way that folk practices have not. Hence, there is a definite payoff for Jackson when he treats questions about the nature of belief as though they were questions about one’s competence with the word ‘belief’.

We won’t push a view on whether it is useful to talk of folk psychology, or our views on persons, or our moral practice, as a collection of sentences in a language. However, suppose it is okay to think in this way. Jackson could retain his talk of theories and his talk of them as logico-linguistic items but ditch the claim that the terms in these theories are words as we ordinarily use them.

The way to do it is to think of the languages of these theories not as languages spoken by the folk or the scientists, but as languages constructed by the theorist–anthropologist–conceptual analyst. Suppose I am interested in finding out what beliefs are. Well, according to Jackson, my first step is to find out what we mean by the word ‘belief’, which might be a bigger job than that because we probably need to understand the whole of folk psychology and hence the network of sentences that give ‘belief’
its meaning. Anyway, we are supposed to get out of this some idea of what it would take for a thing to be a belief. We have claimed that what Jackson should be saying is different. He should be saying that my job is to analyze all our belief-related practices. This more inclusive task will involve discovering the conditions under which one can rightly ascribe beliefs to people and the situations in which the word ‘belief’ can be felicitously uttered, but it will also involve discovering when our actions are best explained by the hypothesis that so-and-so believes that $p$ and not that $q$, and situations where nonlinguistic creatures betray a notion of belief. Having done all of this, I might be able to attribute a folk theory—folk psychology—to the folk (and maybe to the relevant nonlinguistic creatures, if there are any). I will want to be able to express this theory verbally (or at least symbolically) in order to say anything about what it would take to realize the folk-psychological roles, so it makes sense for there to be a language in which I express it. Maybe I will use the word ‘belief’ in that artificial language as a term for the folk-psychological posit that ordinary English-speaking folk are getting at when they too use the word ‘belief’. But it is to be understood that there might be more packed into the semantics of ‘belief’ in the language of my theory (my reconstruction of folk psychology) than is packed into the semantics of ‘belief’ in English. I will have stipulated that my neologism (‘belief’) applies to the practices of non-English speakers and maybe those of certain nonhumans to some extent. This stipulation is not to be abandoned if, as it turns out, English-speakers don’t use ‘belief’ to cover such cases.

Armed with a special vocabulary for my theory, I, the analyst, may blissfully nothing-over-and-aboverize. Good luck to me.

Acknowledgments

Many thanks to Fred Kroon, David Lumsden, Alex Morgan, Tahua O’Leary, and Denis Robinson for their comments on drafts of this essay.

Notes

1. See Child 1993 for such a view and Haugeland 1982 for more cases where both criteria for reduction fail.

2. See Williamson 1995 for a good discussion of this worry. Williamson says, however, that the worry might not arise for “sophisticated” versions of conceptual analysis like that defended by Smith 1994. Williamson would almost certainly include Jackson’s among the suitably sophisticated versions.
3. This approach is inspired by David Lewis’s “An Argument for the Identity Theory” (1966).

4. The connection between conceptual analysis and competence is made explicit in Jackson and Pettit 2002. More indirect evidence comes from Jackson 2004 where the account of semantic competence with referring terms mirrors the story in Jackson 1998a about conceptual analysis.

5. We owe this point and the example types to Denis Robinson: see Robinson 2004.
1 Introduction

The so-called Canberra Plan is a grandchild of the Ramsey–Carnap treatment of theoretical terms. In its original form, the Ramsey–Carnap approach provided a method for analyzing the meaning of scientific terms, such as ‘electron’, ‘gene’ and ‘quark’—terms whose meanings could plausibly be delineated by their roles within scientific theories. But in the hands of David Lewis (1970, 1972), the original approach begat a more ambitious descendant, generalized and extended in two distinct ways: first, Lewis applied the technique to analyze the meaning of terms introduced not just by explicit scientific theories, but also by implicit folk theories such as folk psychology; second, he supplemented the theory to provide an account of the way in which the referents of the analyzed terms might be identified on the basis of empirical investigation.

In the hands of the Canberra Planners, the Ramsey–Carnap–Lewis technique has been generalized further still. As Lewis originally applied the model, theoretical terms were defined in terms of causal roles, and this played a crucial part in determining the kinds of entities that could be identified as the referents of the terms. But Frank Jackson 1998a and others (Tooley 1987; Menzies 1996) have extended the model to apply to terms for entities—such as moral properties and the causal relation itself—that appear to lack causal roles. Instead, practitioners of the Canberra Plan have framed definitions using a more amorphous notion of a functional role within a theory (where ‘functional’ need not mean ‘causal-functional’, apparently). And Jackson has put this generalized model in the service of an ambitious physicalist metaphysics that attempts to show how all properties, including moral properties, can be identified with physical properties.
In this essay, our interest is in the role and place of semantic notions such as reference, satisfaction, and truth in the Canberra program, and in its distinguished Lewisian parent. These questions are of considerable significance in their own right. For one thing, as we shall see, notions such as reference and satisfaction appear to play crucial roles in Lewis’s development of the Ramsey–Carnap technique. However, it is arguable that these uses are eliminable, in a sense familiar from discussions of the pros and cons of deflationary approaches to truth and reference: if so, then Lewis’s own program is compatible with semantic deflationism, at least at this point. In view of the importance of Lewis’s program, and the popularity of deflationism, this is a significant conclusion.

The same question can be raised with respect to the more ambitious program of the Canberra Plan, but in this case, as we want to explain, it acquires new bite. For one thing, it is arguable that the “globalization” of Lewis’s technique envisaged by Jackson and his coworkers requires that semantic notions play the role played by causation in Lewis’s original. If so, it seems to make the global program incompatible with semantic deflationism—which, canonically, wants to deny that semantic notions play such a substantial theoretical role. Evidently, that conclusion, too, is interesting in its own right.

Perhaps more importantly, however, the conclusion seems to leave the Canberra program vulnerable to two kinds of objections. The first is that resting on semantic foundations has the effect of placing the desired metaphysical conclusions out of reach, because there is no prospect of our achieving the kind of knowledge of the relevant semantic relations on which such conclusions would therefore have to rely. The second raises a spectre of circularity, or perhaps incompleteness, in virtue of the global ambitions of the Canberra program—if the program does depend on semantic foundations in this way, can it consistently be applied to the metaphysics of the semantic notions themselves? (Note that Jackson himself takes it that the program is applicable to these notions—indeed, he offers them as the first of his examples of its intended application [Jackson 1998a, 2]; so there are ad hominem grounds for raising the issue, as well as the theoretical grounds already mentioned.)

Our plan of attack is as follows. In section 2 we outline Lewis’s program for theoretical definition and identification, with two issues particularly in view. The first is the question as to whether the program is compatible with semantic deflationism; we argue that it is. The second concerns the precise role of causal notions in the program, and their sensitivity to issues about the precise objectives of the program.
In section 3 we outline the Canberra Plan’s proposed generalization of Lewis’s program. We make good the claim that at least on the most obvious understanding of the goals of the program in question, it requires that semantic notions take over a substantial role played by causal notions in Lewis’s original program, and hence is incompatible with semantic deflationism. This leads us, in section 4, to a discussion of the difficulties just mentioned: the issue as to whether the Canberra program can consistently be applied to the semantic notions themselves, and the question as to whether semantic notions provide a useful route to the investigation of the metaphysics of other topics.

In the brief concluding section 5, we reassess the claims of the Canberra Plan to be an heir to Lewis’s program. We note that our discussion reveals that there are in fact two competing claimants to the Lewisian mantle, in the form of two distinct interpretations of the Canberra program. One version relies on substantial semantic notions, and is accordingly subject to the difficulties we have identified; the other does not rely on such notions, but is correspondingly less ambitious than Lewis’s original program, in a sense our discussion in the earlier sections of the essay will have made clear. In different ways, both options offer us significantly less than the Canberra Plan might have seemed to promise. Our main conclusion is that the choice cannot be avoided: Canberra Planners cannot have their cake and eat it too.

2 Lewis’s Model

Lewis’s model for theoretical definition and identification involves two distinct techniques or stages. It will be important in what follows that we be able to distinguish these stages, so we introduce them separately.

2.1 The First Stage

The first stage or technique of Lewis’s program is the Ramsey–Carnap–Lewis account of the meaning of theoretical terms. On this account, a theory is thought of as providing an implicit functional definition of the terms it introduces. Lewis (1970, 1972) gives an elegant schematic characterization of how to make this definition explicit. Suppose we have a theory, T, that introduces some new terms t₁, . . . , tₙ. These are the T-terms—the theoretical terms. The other terms are the O-terms—the old or original terms whose meaning and reference are understood prior to the introduction of the theory. The theory T can be presented in the form of a single conjunctive sentence—the postulate of the theory. Lewis writes: “It
says of the entities—states, magnitudes, species, or whatever—named by the T-terms that they occupy certain *causal roles*: that they stand in specified causal (and other) relations to entities named by O-terms, and to one another” (Lewis 1972 [in 1999], 254).

Thus the postulate is written:

\[ T[t] \]

By replacing all the terms in the postulate with variables \( x_1, \ldots, x_n \) and prefixing this formula with existential quantifiers, we obtain the Ramsey sentence of T:

\[ \exists x T[x]. \]

This says that there is an \( n \)-tuple of entities satisfying the postulate, or in other words that there is a realization of the theory T. We can also obtain the modified Ramsey sentence, which says that T has a unique realization:

\[ \exists x T[x]. \]

Lewis suggests that if we want a meaning postulate for T, we should adopt what he calls the modified Carnap sentence:

\[ \exists x T[x] \supset T[t]. \]

This says that if T is uniquely realized, the T-terms name the components of this realization. This meaning postulate implies a sentence that explicitly defines the T-terms by means of the O-terms:

\[ t = \text{the unique } x \text{ such that } T[x]. \]

Lewis calls this a *functional* definition: the T-terms have been defined as the occupants of the causal role specified by the theory T—they are the entities, whatever those may be, that bear the specified causal relations to one another and to the referents of the O-terms.

### 2.2 Does the First Stage Require Inflationary Semantics?

We have just employed the phrase ‘the referents of the O-terms’ and (quoting from Lewis) the phrase ‘entities named by the O-terms’. Question: Can these uses of the semantic notions of reference and naming be understood in a deflationary or “disquotational” spirit, or do they indicate that the first stage of Lewis’s program is already committed to employing more robust semantic notions? The answer is that the uses in question can indeed be taken in a deflationary spirit. They meet a need that arises, in effect, simply from the logical generality of the exposition given above.
In any actual case, we can replace an expression such as ‘the referents of the O-terms’ by appropriate uses of the O-terms in question themselves.

Consider, for example, Lewis’s famous application of this account in the service of a functional definition of terms for mental states. He asks us to think of folk psychology as a term-introducing theory, consisting in platitudes regarding the causal relations of mental states, sensory stimuli, and behavioral responses. The general form of these platitudes will be:

When someone is in so-and-so combination of mental states and receives sensory stimuli of so-and-so kind, he tends with so-and-so probability to be caused thereby to go into so-and-so mental states and produce so-and-so behavioral response.

The postulate of the theory is the long conjunction of these platitudes. The T-terms are ‘belief’, ‘desire’, and the like, while the O-terms include ‘sensory stimulus’ and ‘behavioral response’, and more specific terms in the same families, as well as the causal vocabulary itself. On the basis of the postulate, we can form a functional definition of mental states that defines the mental states collectively in terms of their causal relations to stimuli, responses, and each other. And because we can use the O-terms themselves in formulating this definition, we don’t need to employ semantic notions such as reference (at least at this first stage in the program).

A simpler example, to reinforce this point: suppose we are interested in the theory of smoke detectors. Presumably, all smoke detectors have an on-state, which, in a properly functioning detector, is typically caused by the presence of smoke, and causes the emission of a loud noise (or other alarm signal). We could say:

The referent of the term ‘on-state’ is the state that is typically caused by the referent of the term ‘smoke’ and that typically causes the referent of the term ‘loud noise’.

But clearly the reference to referents is otiose: we just told you what the on-state is, without mentioning referents.

In general, then, the first stage of Lewis’s model—the Ramsey–Carnap–Lewis technique for functionally defining theoretical terms—does not make essential use of nondeflationary semantic notions. (It might do so in particular cases, of course: nondeflationary semantic terms might be among the O-terms.)
2.3 The Second Stage
The second part of the model consists in Lewis’s technique for identifying the referents of functionally defined terms on the basis of empirical information. Lewis observed that theoretical identifications such as the identifications of water with H₂O and of light with electromagnetic radiation had previously been thought of as “pieces of voluntary theorizing”: they were hypothesized as bridge laws identifying the entities of one theory with the entities of another theory. But the Ramsey–Carnap–Lewis account of theoretical definition made possible another model of theoretical identification, according to which they are logically implied by the functional definitions of the theoretical entities, taken in conjunction with other bodies of knowledge.

Again, Lewis’s treatment of mental states provides a useful illustration of this technique (and of a further distinction that will play some role in what follows). According to Lewis, the identity of mental state-types with particular physical brain state-types could in principle be established by a simple argument from two premises:

Mental state M = the occupant of the M-causal role R.
Neural state N = the occupant of the M-causal role R.
Therefore, mental state M = neural state N.

Lewis says that the first premise is an a priori truth, supplied by the functional definition of mental states in terms of their causal role. The second premise would be an a posteriori truth, supplied by the advance of neurophysiology. The core of the second stage of Lewis’s program is that what the first stage provides, in effect, is a nontrivial target for empirical investigation: in this case, investigation of what it is, in fact, that plays the causal role R.

2.4 The Role of Causation
Note that there is always a trivial answer to the above question, namely, that it is precisely mental state M that plays causal role R. The second stage of Lewis’s program can be construed as offering us a guarantee that there is a nontrivial answer, and a prescription for finding it. We are interested in highlighting the role that causation plays in underwriting this guarantee. To this end, it will be helpful to step back a little, conceptually and historically, and consider the original incarnation of Lewis’s proposal from “An Argument for the Identity Theory” (Lewis 1966). Here, as the title of Lewis’s paper indicates, the emphasis is not so much on the particular identification of mental states with brain states, or even on the method
for finding such an identification, but on the guarantee that there is some such identification to be found: in other words, on the general argument for physicalism about the mental.

The structure of Lewis’s paper makes it explicit that his argument for materialism about mental states has two premises. And as he says, “[t]he first of my two premises for establishing the identity theory is the principle that the definitive characteristic of any experience as such is its causal role” (Lewis 1966, 19). In other words, the first premise is simply what later comes to be formalized in terms of the Ramsey–Carnap technique. Lewis goes on to emphasize (ibid., 17) that this premise is not in itself a materialist premise: on what experiences actually are, it is neutral, so long as they play the required causal roles.

As for the second premise, Lewis characterizes it as “the explanatory adequacy of physics” (Lewis 1966, 23). It is the principle that “there is some unified body of scientific theory, of the sort we now accept, which together provide a true and exhaustive account of all physical phenomena” (ibid.). As Lewis goes on to explain, to assume this principle is not to assume physicalism itself: “My second premise does not rule out the existence of nonphysical phenomena; it is not an ontological thesis in its own right. It only denies that we need ever explain physical phenomena by nonphysical ones” (ibid., 23–24). The crucial point is that in the light of the first premise, “none of these nonphysical phenomena can be experiences.” Why? Because “they must be entirely inefficacious with respect to all physical phenomena,” whereas the first premise tells us that experiences are not inefficacious in this way: on the contrary, they have physical causes and effects, such as bodily stimuli and behavior.

In general, then, one thing that the second stage of Lewis’s program can be taken to provide is an argument for physicalism about the domain in question, constructed by analogy with the mental case. We stress that this kind of application of the program does not presuppose physicalism. Rather, it argues for physicalism, by invoking two premises: first, the claim that the entities in question are characterizable in terms of their causal roles; and second, what amounts to a physicalist principle about causation itself, to the effect that nonphysical things do not have physical effects. If the program is to be applied in this way, then, causation plays an essential role—it is at the core of the crucial second premise of the argument—which means, in turn, that it is vital that the theoretical roles identified at the first stage of the Lewis program be causal roles. (Later we will ask whether the Canberra generalization of Lewis’s program claims an analogous second stage, and if so, what plays the role that causation plays in Lewis’s program.)
The second stage of Lewis’s program can also be taken in a less general way. As we noted above, the identification of a state M as the occupant of some role R does not in itself yield a nontrivial identification—after all, the best answer to the question ‘What plays the R role?’ might be simply, ‘Why, M, of course!’ The second stage of Lewis’ program offers us something more than this trivial answer. Because the first stage of the program identifies the entities in question (originally, mental entities) as occupants of (physically efficacious) causal roles, and the second premise assures us that the study of such causal roles lies within the scope of the physical sciences, their combination gives us a route to a nontrivial identification, at least in principle: roughly, it tells us that physics will get us there.

Thus, to summarize, we have distinguished two versions of the goal of the second stage of Lewis’s program. The first and more general version offers an argument for physicalism about the domain in question. The second version offers what we might call a particular technique for theoretical identifications: it instructs us to investigate the nature of the occupants of those causal roles delineated in the first stage of the program. The second relies on the first, in the sense that it is the general argument for physicalism, or at least the causal closure principle on which that argument relies, that guarantees that the methods required are essentially those of the physical sciences in general.

We have noted that causal notions play an absolutely central role in both versions. Without an appeal to causation, there would be no general second stage to Lewis’s program. The first stage could stand alone, of course, as an account of the meaning of the T-terms in question, but it would not yield an argument for physicalism, and it would not provide any general method for making theoretical identifications. (At least, it would not yield a method beyond the obvious one, viz., the recommendation to look for the x such that T[x], where \( \exists x T[x] \) is the Ramsey sentence in question.)

It will be helpful to have a name for the role that causation plays here. Since it is providing a hook, or tag, in terms of which these issues of identity can be addressed by the second stage of Lewis’s program, we shall say that it is an ID tag. Later, we will be interested in the question as to whether semantic notions are required as ID tags, in the globalized version of Lewis’s program.

3 To the Canberra Station: The Semantic Route from Lewis to Jackson

As we have already noted, the essence of the Canberra Plan consists in an extension of Lewis’s program to a range of cases not envisaged by Lewis.
himself. For example, Frank Jackson (1998a) invokes Lewis’s model to provide an account of the nature of moral properties. Again, the program has two stages. First, Jackson asks us to consider how current folk morality, as reflected in our intuitions about how descriptive and moral terms are interconnected, might develop into a mature folk morality in the limit of critical reflection. The key postulate of mature folk morality will then consist of a conjunction of all the platitudes that describe the relationships between nonmoral descriptions of situation and moral descriptions, the interconnections between moral descriptions, and the relationship of moral judgments to motivation and behavior. If one introduces variables for the names of all moral properties and binds them with quantifiers, one obtains the modified Ramsey sentence. Then, by adopting the modified Carnap sentence as a meaning postulate for mature folk morality, one can formulate definitions of moral terms by reference to their functional role in this theory.

Jackson’s ambitions are not limited to this first-stage application of Lewis’s model, however. On the contrary, he argues that his doctrine of “moral functionalism” lends itself to theoretical identification of ethical properties with descriptive ones. He considers, for example, the possibility that the best solution to the equations of mature folk morality might be ones that yield a posteriori identifications of rightness with maximizing expected hedonic value and goodness with positive expected hedonic value (Jackson 1998a, 142).

However, as Jackson himself notes, the functional roles of moral properties are not typically causal roles. The platitudes of mature folk morality are rarely couched in causal terms, apparently. For example, Jackson writes, one platitude might be that a fair division of some good is, other things being equal, morally better than an unfair division—but that does not mean that being fair causes things to be morally better (ibid., 131).

In the light of this fact—the fact that in this case the relevant functional roles are not causal roles—we want to ask the following question: What plays the general methodological role in this case that causation plays in Lewis’s less ambitious program, namely, as we put it, the role of an ID tag?

One more example, before we return to this issue. Once one allows that the technique for giving functional definitions need not appeal to causal roles, one must allow that the technique may be applied to the causal relation itself. Indeed, several writers, including one of the present authors, have explored this possibility (Menzies 1996; Tooley 1987). Thus, if one thinks that we have a folk theory of causation that implicitly treats causal
terms as theoretical terms, one might apply the Ramsey–Carnap–Lewis technique in the familiar way to define them explicitly. One such application might yield the definition that causation is the intrinsic relation that typically accompanies counterfactual dependence between distinct events (Menzies 1996). It might be hoped that such a definition could also be harnessed to provide an a posteriori identification of causation with some other physically specifiable relation, say energy-momentum transfer. But again, the question arises as to what ID tag could mediate this identification.

3.1 The Search for ID Tags
Once again, it is important to emphasize that like its Lewisian parent, the Canberra version of the Ramsey–Carnap–Lewis technique can be deployed in ways that are more or less ambitious, metaphysically speaking. The metaphysically unambitious deployment utilizes only the first stage of the technique, in the service of a neutral account of the meaning of theoretical terms. As in the case of the Lewisian version, it is evident that this stage does not require substantial semantic notions—semantic references can be understood in a deflationary spirit.

Our concern is with the metaphysically ambitious deployment of the generalized program—the deployment that counts theoretical identification among its ambitions (in either the more or the less general senses, as we distinguished them in section 2.3—the general argument for physicalism concerning the area in question, or the particular prescription for a posteriori identifications). We saw that in the Lewisian case, this deployment relies on causation as an ID tag. Hence our question—What, if anything, can play this role in the ambitious version of the generalized program?

It seems to us that there is only one possible answer to this question. Causation gets replaced by one or more semantic notions, such as reference or satisfaction: semantic notions come to provide the ID tags, providing the basis for a general argument for physicalism and the mediating links for particular a posteriori identifications. For example, let us suppose that we wish to produce a completely general argument for physicalism, without recourse to causal ID tags. Then it seems that the only candidate for a completely general argument would be one like the following, which makes essential use of semantic ID tags:

t is the referent of the term ‘the unique $x$ such that $T[x]$’;
All referents are physical entities;
Therefore, $t$ is a physical entity.
Here the first premise uses the Ramsey–Carnap–Lewis technique to define the meaning of the term ‘t’ by way of its role in a relevant theory T, with the explicit addition of a semantic tag. The second premise is a general premise that plays the role in this argument that the physical causal closure principle played in Lewis’s original argument for physicalism. As we saw in section 2.4, the causal closure principle is able to act as a premise in an argument for physicalism because it does not presuppose physicalism: it leaves open the possibility, for example, that there are non-physical entities that do not have any effect on the physical. Correspondingly, the second premise in the argument above does not itself presuppose the physicalist conclusion it purports to establish: it leaves open the possibility that there are nonphysical entities that are not the referents of any theoretical terms. Nonetheless, the principle is sufficiently general that it can support the conclusion that anything referred to by a theoretical term is a physical entity.

As we say, this seems to be the only kind of argument available to an advocate of the Canberra Plan who wishes to generalize Lewis’s original argument for physicalism. Unless semantic ID tags are to replace causal ID tags, there seems to be no prospect of a general analogue of the crucial second premise of Lewis’s argument. (Canberra Planners might well invoke some other argument for physicalism, of course. Our point is simply that without a suitable ID tag to replace causation, such an argument cannot be an analogue of Lewis’s argument for materialism.)

Thus if semantic notions did not play the role of ID tags in the Canberra Plan’s extended Lewisian program, the Plan could not provide an analogue of what we called the general version of the second stage of Lewis’s own program: a schema for an argument for physicalism, applicable wherever the program itself is applicable. This leaves the question as to whether the program could provide an interesting analogue of the second version of the second stage of Lewis’s own program, namely, a schema for a method of a posteriori identifications. Here, Lewis’s principle of the causal closure of the physical world offered us the advice, in effect, that we should always look to physics to find the occupants of the relevant causal roles—and again, the generality of the prescription rests on the fact that causation is playing the role of an ID tag. If there is to be no general substitute for causation as an ID tag in the extended version of the Lewisian program, there can be no general prescription of this kind. Instead, we are left with something like this:

To find out what t is, write down what you know about t in the form ‘t is the unique x such that T[x]’; and then ask yourself what is the thing x such that T[x].
There is no doubt that this is a general prescription of some kind. What is questionable is whether it provides us with anything new or nontrivial, or at any rate, anything whose novel or nontrivial elements amount to more than the application of the formal Ramsey–Carnap technique itself—which, as we have stressed, belongs to the uncontroversial first stage of the Lewisian and Canberra programs.

Thus we conclude that the unambitious, one-stage version of the Canberra program does not require robust, nondeflationary semantic notions; but the second stage does require such notions, to play the role of ID tags, if the program is to offer what Lewis's program offers at this second stage: a general schematic argument for physicalism, and a non-trivial methodology for theoretical identification, on a case-by-case basis.

With these conclusions in hand, we now turn to their consequences. As we have already noted, there are two main questions to consider. The first is whether the Canberra program can be applied to the metaphysics of the semantic relations themselves (or whether, on the contrary, the role that semantic notions are required to play as ID tags stands in the way of the application of the Canberra technique to these notions themselves). The second is whether there is some general difficulty in relying on semantic notions in this way—a difficulty that would be problematic in general, and not merely in the case of any proposed application to the case of the semantic notions themselves.

4 Semantics on the Canberra Plan?

4.1 The Local Case

We have seen in the last section that the application of the Ramsey–Carnap–Lewis technique to a theory presupposes that there is a division between the T-terms introduced by the theory and the O-terms understood before the theory’s introduction. The technique relies on such a division because the T-terms are defined in terms of O-terms, taken in conjunction with the logical vocabulary. We have also seen that the set of T-terms and O-terms vary from one theory to another. The term for the causal relation may be an O-term in the theory of folk psychology, but it is a T-term in the folk theory of causation.

We have also seen how the practitioners of the Canberra Plan have progressively extended the technique to encompass a broader range of terms beyond those mentioned in explicit scientific theories. Following Lewis's application of the technique to folk psychology, they have applied it to folk theories of colors, causation, and ethics. Jackson also suggests applying
it to the case of the semantic properties themselves: indeed, this is his very first example of the kind of “location problem” to which he takes the methodology to be applicable, in Jackson 1998a. We are now interested in the status of such an application, in the light of our conclusions in the previous section. (For simplicity we assume that the semantic notions stand or fall together in this respect—in practice, clearly, there is scope for defining some of them in terms of others, but at some point, our issue will arise for whichever notion or notions are treated as basic.)

The key issue now is whether semantic notions can function as ID tags, in a generalized Lewisian approach to the identification of semantic properties and relations. And the answer, fairly obviously, is that they cannot. Why? Simply because the method requires the elimination of the target T-vocabulary, which would not be possible if the target semantic terms are themselves part of the O-vocabulary.

This point has perhaps been obscured by the naturalness of a harmless (because eliminable) use of disquotational semantic vocabulary. Thus if $\exists x \text{Sat}[x]$ is the Ramsey sentence for the T-term ‘satisfaction’, we may say harmlessly that satisfaction is the unique $x$ (if such there be) such that $x$ satisfies $\text{Sat}[x]$. The reason this is harmless is that the italicized use of ‘satisfies’ is entirely eliminable. We may say, equivalently (if less elegantly), that satisfaction is the unique $x$ (if such there be) such $\text{Sat}[x]$. Once the first stage of the application of the Ramsey–Carnap–Lewis technique to “satisfaction” is characterized in this latter way, it is apparent that there is no scope for the notion of satisfaction itself to function as an ID tag—to provide the necessary “hook” for a further a posteriori identification of the satisfaction relation.

Note that this doesn’t imply that there can be no metaphysics of satisfaction; only that such a metaphysics cannot avail itself of the generalized Lewis–Jackson methodology, if this is conceived as invoking semantic relations for the job for which Lewis’s program invoked causal relations. We still have the “thin” version of the program, the kind we mentioned above. This doesn’t offer a new technique for the metaphysics of semantic properties (apart from the first-stage Ramsey–Carnap technique itself); but it isn’t incompatible with old techniques.

4.2 The Global Case
The second question we raised at the end of section 3 is whether there is any general difficulty in relying on semantic notions as ID tags in the generalized Lewis program—any difficulty that might be problematic in general, and not merely for the proposed application to the case of the semantic notions themselves. We want to raise two points of this kind.
4.2.1 Stich’s Problem  For the first problem, we turn to an excellent discussion of a closely related issue by Stephen Stich, in the first chapter of his *Deconstructing the Mind* (Stich 1996). Stich is concerned with eliminativism about folk-psychological notions such as belief and desire. He notes that many philosophers take the eliminativist thesis to be the view that the terms ‘belief’ and ‘desire’ do not refer. But if that is how eliminativism is to be characterized, Stich argues, then in order to assess it we need a theory of reference—a theory capable of guiding our judgment about whether these terms do succeed in referring.

Stich argues that this leaves metaphysics hostage to the inevitable indeterminacies in a scientific theory of reference. In other words, it means that we can’t decide whether eliminativism is true until we sort out the issue between competing theories of reference—and that’s likely to mean “never,” given the nature of scientific theory. (The threat of deflationism also lurks in the background here, of course, but we leave that aside.) Even worse, it would seem that in crucial cases, the metaphysics needs to precede the theory of reference. In order to decide what relation reference is, presumably, we need to be able to examine typical cases. In other words, we need to be able to study the various relationships that obtain between words or thoughts on the one side, and the items to which they (supposedly) refer on the other. But how can we do this in the case of ‘belief’ and ‘desire’, while it is in doubt whether these terms refer to anything? In order to know where to look, we’d have to know not only that they refer, but also to what.

Thus we have two problems for eliminativism, if it is to rely on semantic relations such as reference: the referential indeterminacy problem and the precedence problem, as we might call them. Clearly, both are problems not simply for eliminativism, but for any metaphysical view that relies on reference in this way. Both problems apply just as much if the question is “What is belief?,” if this is to be understood as ‘To what does the term ‘belief’ refer?’, as they do to the question “Are there beliefs?,” understood as ‘Does the term ‘belief’ refer to anything?’ Thus they apply in particular to any version of the generalized Lewisian program that seeks to employ reference as an ID tag.

Stich’s own response to the problem is to abandon semantics, and ask the relevant metaphysical questions in material form: ‘Are there beliefs?’, in place of ‘Does the term ‘belief’ refer to anything?’, for example. This certainly seems the appropriate response in some cases (folk psychology might be more controversial than Stich thinks, perhaps, but chemistry isn’t, for example). But a Canberra Planner could only follow Stich down this path at the cost of abandoning the ambitious program in which (as
we explained in section 3) semantic notions are required to play the role that causal notions play in Lewis’s original program.

In summary, then, the first general problem with relying on semantic notions as ID tags is that these notions are not capable of bearing the weight, in at least two senses:

1 Referential indeterminacy  There is no realistic prospect that a theory of semantic notions will ever be sufficiently well founded, or sufficiently uncontroversial, to provide what causal notions arguably do provide in the original Lewisian cases, namely, a practical basis for a posteriori investigation of identity questions.

2 Priority One source of the indeterminacy problem is that in the cases that matter, our knowledge of the relevant facts about semantic matters is inevitably subordinate to our knowledge of the relevant metaphysical matters. It is crying for the moon to suppose that we might reverse things, and use semantics as our guide to metaphysics.

4.2.2 Circularity Problems  The second class of problems for relying on semantic notions in the role of ID tags in the generalized Lewis program turn on the fact that (as we have already observed) these notions cannot play such a role in their own case. In other words, a version of the program founded on these semantic notions cannot turn its own spotlight on the semantic notions themselves. (This was a simple logical point: if the semantic terms are among the T-terms, they are not available as ID tags.) In our view, there are at least two ways to argue that this consequence is unsatisfactory.

The first is a methodological point. Interpreted in such a way as to rely on semantic notions as ID tags, the Canberra program is offering us a conception of the task of metaphysics: the task, in effect, is to investigate the referents of our words and thoughts, after due regimentation by Ramsey–Carnap methods. However, we have seen that this task is incoherent in the case of the semantic notions themselves. Either, therefore, the conception of the task of metaphysics is flawed, or metaphysics is essentially uncompletable—inapplicable just where it matters most, in fact, given the role of the semantic notions in grounding the entire program.

The second argument attempts to give the first a little more bite. It points out that in effect, this version of the generalized program offers us a semantic criterion for realism about any metaphysical matter: to be a realist about a domain is to believe that the terms characteristic of the domain in question do succeed in referring. A little more generally, the approach gives us
a semantically characterized account of *what the issues are* for what Jackson calls “serious” metaphysics. Concerning any target subject matter, the key issues are *whether* the terms characteristic of that subject matter succeed in referring, and if so, *to what*. The problem is that this account is not applicable in the case of the semantic notions themselves. We can adopt some other account of the metaphysical issue in the case of the semantic notions, of course—as we noted above, we can simply ask whether such relations exist, and if so, what they are. But this involves turning away from the semantically characterized criterion we have been offered for other cases.

One manifestation of this difficulty is the incoherence of irrealism about semantic properties, if irrealism is understood in the “failure of reference” way. This problem is noted by Boghossian 1990, for example. *Pace* Boghossian, however, it does not provide a transcendental argument for realism about semantic properties. For it doesn’t exclude other kinds of irrealism about semantic properties, such as the materially constituted view that there are no such things, or simply the deflationist view that there are no substantial semantic relations. (What is incoherent, in each case, is going on to try to express these forms of irrealism in the semantic fashion.)

### 4.3 The Best-Case Scenario?

These arguments leave space for one “strong” version of the Canberra program. This version would accept that the program relies on the semantic notions as ID tags, and therefore that it isn’t applicable to the semantic notions themselves; but it would claim that it is nevertheless an advantage to be able to put all our other metaphysical eggs in the semantic basket, as it were—to conduct the metaphysics of all other topics in semantically mediated vocabulary. This view is not incoherent, in our view, and would have a legitimate claim to be a descendant of the original Lewisian program, in its strong form. However, there seems little prospect that it could yield dividends comparable to those of the original Lewisian program, for the reasons we have noted. (In essence, they are the reasons identified by Stich: whatever the merits of Lewis’s injunction to investigate causes via physics, investigating reference relations is a much less promising project.) And it could not claim the attractions of being a *global* program for metaphysics, because it cannot apply to its own foundations.

### 5 Conclusion

We have argued that the alternative to this strong version of the Canberra program is a version significantly weaker than Lewis’s original: weaker
precisely in the sense that in lacking any general alternative to causation to play the role of an ID tag, it is necessarily less ambitious at the second stage of Lewis’s program—at the stage at which we move from conceptual analysis to metaphysics, in effect. Here, Lewis’s causal closure principle offers a general schematic argument for physicalism about any domain to which his version of the program is applicable, and a general technique for a posteriori identifications. Without an alternative ID tag to call its own, the Canberra program is inevitably weaker at this point.

Thus we conclude that the lineage from Lewis’s program to its Canberra descendants actually leads to a fork. Follow one branch, and we reach a robust but somewhat unappealing metaphysical descendant, irredeemably dependent on semantic foundations that are inaccessible both to its own methods and to useful a posteriori investigation. Follow the other branch, and we reach a different descendant—healthier than its sibling, to be sure, in not being dependent on something that lies forever out of reach, but weaker and less ambitious than its famous parent. We will not express an opinion as to which has the better claim to be Lewis’s true heir, but we want to insist that they cannot share the mantle.

Note

1. It is arguable that Lewis himself ought to count as a practitioner of the Canberra Plan by these lights, on the grounds that his mature view does not seem to restrict the relevant theoretical roles to causal roles. If so, then what we say below about the Canberra Plan applies also to the mature version of Lewis’s view. But our interest is in some contrasts between the causal case and the general case, and for terminological purposes it will be convenient to continue to refer to the former as the Lewisian view.
II Metaphysics
9 Ramseyan Humility

David Lewis

1 The Thesis of Humility

Rae Langton has recently argued that Kant is not the thoroughgoing idealist we mistook him for. Our irremediable ignorance of noumena is not ignorance of some remarkable things that are hidden from us behind a veil of appearances. Rather, it is our ignorance of the intrinsic properties of substances. The substances that bear these intrinsic properties are the very same unhidden substances that do indeed affect us perceptually. But they affect us, and they affect other things that in turn affect us, in virtue of their causal powers, which are among their relational properties. Thereby we find out about these substances as bearers of causal powers, but we find out nothing about them as they are in themselves (Langton 1998).

Langton makes an impressive case for her interpretation. But in doing so, she disagrees extensively with previous commentators—who, in turn, disagree extensively with one another. I am not entitled to an opinion about these disputed exegetical questions. In any case, my interest is not in whether the thesis of Humility, as she conceives it, is Kantian, but rather in whether it is true. I shall argue that it is, or at least that something very like it is.

Many of us nowadays believe that things do indeed affect other things in virtue of their intrinsic properties. Something is disposed under certain circumstances to cause a response in something else because it has some intrinsic ground which, in accordance with the contingent laws of nature, would under those circumstances cause the response. If that is true, it might seem to demolish the case for Humility. Langton disagrees:

[If] the ground is distinct from the power, and contingently connected with it, then our orthodoxy is faced with a conclusion surprisingly similar to Kantian Humility. . . . [Our name for the ground] becomes the name for a something-we-
know-not-what—ominously similar to a Kantian thing in itself. Our contemporary orthodoxy must concede that it, too, is faced with a conclusion similar to Kant’s—that there are intrinsic features of the world with which we can never become acquainted. (Langton 1998, 176)

To be the ground of a disposition is to occupy a role, but it is one thing to know that a role is occupied, another thing to know what occupies it.

The point generalizes. Being the ground of a certain disposition is only one case among many of role-occupancy. There are a variety of occupied roles, among them nomological roles and others as well. Quite generally, to the extent that we know of the properties of things only as role-occupants, we have not yet identified those properties. No amount of knowledge about what roles are occupied will tell us which properties occupy which roles. I do not agree with Langton that the predicament is “ominous” but I do agree that we are in it. That is what I shall argue here. First I shall argue from fairly weak assumptions that Humility applies to at least some of the fundamental properties. Next I shall argue from somewhat more contentious assumptions that it applies to all of them—or if it doesn’t, we are in no position to know that it doesn’t, which is no less of a setback for our aspirations to knowledge. Finally I shall argue that if Humility applies to most or all of the fundamental properties, then it spreads to a great range of less than fundamental properties, intrinsic and extrinsic alike.

2 Fundamental Properties

Fundamental properties are those properties that I have elsewhere called ‘perfectly natural’. They are not at all disjunctive, or determinable, or negative. They render their instances perfectly similar in some respect. They are intrinsic; and all other intrinsic properties supervene on them. They are not conjunctive or structural.

Unfortunately, all that I’ve just said about fundamental properties would, if analyzed, lead us back in a definitional circle to the notion of a fundamental property. If we had no initial understanding of such notions as ‘disjunctive’, ‘determinable’, and so on, we would gain no understanding by running around the circle. If, on the other hand, we do have some initial understanding—and I think we do, as witness our abilities to classify—then the circle is benign. When we run around it, our initial understandings of the several notions we meet reinforce one another.

Properties are abundant, as abundant as classes of possibilia (which is what I take them to be). But the great majority of properties are utterly
miscellaneous disjunctions, not at all natural. Properties that are even somewhat natural are sparse, and perfectly natural (that is, fundamental) properties are sparser still. I do not know whether some properties correspond to universals or to tropes, but it offends against Occam’s razor to imagine that all the abundant unnatural properties do. Rather, it is the fundamental properties, or maybe the near-enough fundamental properties, that are plausible candidates for corresponding to universals or tropes.

Fundamental properties figure in a minimal basis on which all else supervenes. No two possible worlds just alike in their patterns of instantiation of fundamental properties could differ in any other way.

I speak of ‘fundamental properties’ for short, but they fall into several categories. There are all-or-nothing monadic properties. There are all-or-nothing n-adic relations, at least for smallish n. There are properties that admit of degree, that is, magnitudes; more generally, there are scalar-valued, vector-valued, tensor-valued, . . . magnitudes. There are relational magnitudes. Maybe my list is too long; maybe the magnitudes could somehow be reduced to all-or-nothing properties and relations, but that is a question I shall not take up here.

Scientific theorizing and the discovery of fundamental properties have gone hand in hand. For instance the discovery of the phenomena of electromagnetism and the laws governing them was inseparable from the discovery of the previously unknown, and very likely fundamental, properties of positive and negative charge. So if we had a true and complete “final theory,” it ought to deliver a true and complete inventory of those fundamental properties that play an active role in the actual workings of nature.5

That inventory might omit two kinds of fundamental properties: idlers and aliens. Idlers are those fundamental properties, if any, that are instantiated within the actual world, but play no active role in the workings of nature. Aliens are those fundamental properties, if any, that are instantiated within unactualized possible worlds but not within the actual world. We shall consider later whether we have any reason to believe, or any reason to disbelieve, in idlers or in aliens.

3 Ramsification

Let T be the true and complete final theory just considered. The language of T contains T-terms: theoretical terms implicitly defined by T. And there is all the rest of our language, call it O-language. ‘O’ stands for ‘old’; it is the
language that is available to us without benefit of the term-introducing theory T. ‘O’ does not stand for ‘observation’. O-language is not meant to be a ‘pure observation language’, and indeed I doubt that there could be any such thing. I’ll assume, however, that O-language does suffice to express all possible observations, whatever else it may also be able to do. I’ll assume also that O-language is interpreted—never mind how. I do not assume it to be first-order, or extensional, or finitary, or free of indexicality. Nor do I assume it to be unmetaphysical, suited only to talk of everyday matters.

All fundamental properties, except for idlers and aliens if such there be, are mentioned in the theory T. If T is the limit (perhaps never reached) of a process in which theorizing and the discovery of fundamental properties go hand in hand, then the fundamental properties mentioned in T will be named by T-terms. I assume that no fundamental properties are named in O-language, except as occupants of roles; in which case T will name them over again, and will say that the property named by so-and-so T-term is the occupant of such-and-such role.

I’ll assume, until further notice, that at least two fundamental properties fall in the same category. They might be two monadic properties: unit positive charge and unit negative charge, say. Or they might be two relational magnitudes: space-like distance and time-like distance, say. An inconclusive reason to assume this is the expectation that current physics is not entirely wrong in its inventory of fundamental properties. Fundamental properties belonging to one-membered categories, if such there be, fall outside the scope of my present argument for Humility.

The theory T consists of all the logical consequences of a sentence we shall call the postulate of T. This is not a substantive assumption, since I have not said that the postulate is of finite length; but if it were not, it is hard to see how any of the theorems of T could deserve the name of laws. The postulate can be written as T(t₁, . . . , tₙ), where t₁, . . . , tₙ are the theoretical terms of T, and all other language that appears in the postulate is O-language. We can assume without loss of generality that the T-terms are names, since given suitable copulas, predicates can be replaced by names of properties; and we can assume that no two of them name the same thing. (If they did, a true and complete theory would say so; but then it could be better formulated by using one name twice over.) Some T-terms might be names of other things (for instance the center of the universe, in the unlikely event that it plays a special theoretical role) but for the most part the T-terms will name fundamental properties. When they do, the postulate ought to say so, and we have assumed that O-language is indeed capable of saying so.
Replacing T-terms by variables, we get a formula $T(x_1, \ldots, x_n)$. An $n$-tuple may or may not satisfy this formula, with respect to the actual world and the fixed interpretation of O-language. One which does is called an *actual realization* (for short, a realization) of $T$. One which might is called a *possible realization* of $T$.

We have assumed that a true and complete final theory implicitly defines its theoretical terms. That means that it must have a unique actual realization. Should we worry about symmetries, for instance the symmetry between positive and negative charge? No: even if positive and negative charge were exactly alike in their nomological roles, it would still be true that negative charge is found in the outlying parts of atoms hereabouts, and positive charge is found in the central parts. O-language has the resources to say so, and we may assume that the postulate mentions whatever it takes to break such symmetries. Thus the theoretical roles of positive and negative charge are not purely nomological roles; they are locational roles as well.

Prefixing existential quantifiers to the formula just considered we get the *Ramsey sentence* of $T$: ‘For some $x_1, \ldots, x_n$, $T(x_1, \ldots, x_n)$’. It says that $T$ has at least one actual realization (see Ramsey 1990, chapter 6). What we need to know about the Ramsey sentence is that it logically implies exactly the O-language sentences that are theorems of $T$. Equivalently: exactly those that are logically implied by the postulate of $T$.

O-language, we assumed, is rich enough to express all possible observations. Therefore any predictive success for $T$ is equally a predictive success for the Ramsey sentence of $T$. Since the evidence for $T$ consists in its record of predictive success, there is no way to gain evidence for $T$ that is not equally evidence for the Ramsey sentence.

Though our theory $T$ has a unique actual realization, I shall argue shortly that it has multiple possible realizations. Suppose it does indeed have multiple possible realizations, but only one of them is the actual realization. Then no possible observation can tell us which one is actual, because whichever one is actual, the Ramsey sentence will be true. There is indeed a true contingent proposition about which of the possible realizations is actual, but we can never gain evidence for this proposition, and so can never know it. If there are multiple possible realizations, Humility follows.

Or rather, Humility about some members of the actual realization follows. We have $n$ T-terms, so possible realizations will be $n$-tuples. Suppose all possible realizations of $T$ have the same 12th and 14th members, but differ elsewhere. Then we can gain evidence that will uniquely identify the 12th and 14th members of the actual realization, but not the other members.
How does the theory T go beyond its Ramsey sentence? Exactly by adding the Carnap sentence: the truth-functional conditional with the Ramsey sentence as antecedent and the postulate as consequent (Carnap 1963, section 24 D, 963–966). The two things we need to know about the Carnap sentence are, first, that the Ramsey and Carnap sentences together are logically equivalent to the postulate; and, second, that the Carnap sentence logically implies no O-sentences except for logical truths. The Carnap sentence of a theory constrains the reference of the theoretical terms, and that is all it does. If the theory is uniquely realized, it says that they refer to the corresponding members of the unique realization. If the theory is multiply realized, it says that they refer to the members of some realization, but it doesn’t say which one. Nor does it say what happens if the theory is unrealized. But neither of those silences need concern us here: our true and complete final theory T is uniquely realized.

4 Combinatorialism and Quidditism

Suppose we have the actual realization of T. Maybe some members of the n-tuple that realizes T are not fundamental properties, or maybe some belong to single-membered categories. Hold those ones fixed. Permute the rest within their categories to obtain a new n-tuple. It too would realize T.

Suppose, for instance, that we start with the actual world, and we permute two fundamental monadic properties F1 and F2, these being the actual referents of the T-terms t1 and t2, leaving all else fixed. Exchange F1 and F2 throughout the world. (Suppose for simplicity that possibilities are worlds like ours; though if instead they were some sort of representations of worlds, my argument would go through mutatis mutandis.) Then F2 will be found in exactly those places in space and time (or, more generally, in the pattern of instantiation of fundamental relations and magnitudes) that correspond to the places where F1 was found originally; and vice versa. And the laws of nature governing F1 in the permutation will be just the same as the laws governing F1 originally (more precisely, the laws governing F2 vis-à-vis F1 in the permutation will be the same as those governing F1 vis-à-vis F2 originally); and vice versa. But what could T tell us about besides the locational and nomological roles of the referents of its T-terms? So if this permutation really is a possibility, it is a possible realization of T.

Possibility is governed by a combinatorial principle (see Lewis 1986a, sec. 1.8; Armstrong 1989). We can take apart the distinct elements of a possibility and rearrange them. We can remove some of them altogether.
We can reduplicate some or all of them. We can replace an element of one possibility with an element of another. When we do, since there is no necessary connection between distinct existences, the result will itself be a possibility. How much this means depends on what we take the distinct elements to be. Here, let us take them to include not only spatiotemporal parts, but also abstract parts—specifically, the fundamental properties.

Combinatorialism tells us that the laws of nature are contingent. 12 Let it be a law that every F is a G; combinatorialism generates a possibility in which an F is not a G, so that this law is violated. Combinatorialism therefore collides head on with the view that the laws of nature are necessary. 13 That is the principal reason why some doubt combinatorialism. For myself, I find combinatorialism far more compelling than the alleged necessity of laws, but I shall not try to adjudicate that question here.

Combinatorialism tells us that possibility is preserved under permutations of items—at least if they are items from the same category. If it is possible that –A–B–C–, and if A, B, and C, are, say, all-or-nothing monadic properties, then it is also possible that –C–A–B–. The actual realization of T is a possible realization; we permute items within more-than-one-membered categories; and what we get is also a possible realization of T. But is it different from the one we had before?

Quidditism is the premise that tells us that the permutation is indeed a different possibility (see Black 2000). Two different possibilities can differ just by a permutation of fundamental properties. They do not differ in whether T is realized, or in what we observe. Given combinatorialism and quidditism, our argument for Humility is complete. We have the actual realization of T, permutations thereof yield different possible realizations of T, and we have no way to tell which one is the actual realization.

Quidditism is to properties as haecceitism is to individuals. If we start with a possibility and permute individuals, combinatorialism says that we get a possibility; haecceitism says that it is a different possibility. Haecceitism says that two possibilities can differ just by a permutation of individuals (see Lewis 1986a, sec. 4.4). I accept quidditism. I reject haecceitism. Why the difference?

It is not, I take it, a difference in prima facie plausibility. In both cases alike, haecceitistic or quidditistic distinctions between possibilities seem offhand to make sense. In both cases alike, however, we can feel an uncomfortable sense that we are positing distinctions without differences. To dislike distinctions without differences is to feel some sympathy toward (nontrivial) principles of identity of indiscernibles. To reject haecceitism
is to accept identity of qualitatively indiscernible worlds; to reject quidditism is to accept identity of structurally indiscernible worlds—that is, worlds that differ just by a permutation or replacement of properties. Such principles, applied within a single world, prove too much: for instance, they force us to reject the possibility of a world of two-way eternal recurrence.14 Are they any better if applied to two different worlds?15

In both cases alike, this conflict of intuitions is inconclusive, but it seems to me to favor both haecceitism and quidditism. However, haecceitism leads to trouble in a way that quidditism does not. Suppose we allegedly have two indiscernible worlds that differ haecceitistically: at the place where individual A is in \(W_1\), B is in \(W_2\), and \textit{vice versa}. We ask: in virtue of what does B rather than A occupy that place in \(W_2\)? Because the worlds are indiscernible, it seems we can have no informative answer. We can only say “It just does”. That is, we need to invoke the transworld identity of B in \(W_1\) with B in \(W_2\).

Similarly, \textit{mutatis mutandis}, if we have two structurally indiscernible worlds that differ quidditistically: where property F is in \(W_1\), G is in \(W_2\) and \textit{vice versa}. In virtue of what does G rather than F occupy that place in \(W_2\)? It just does. That is to say, we invoke the transworld identity of G in \(W_1\) with G in \(W_2\). Unmysterious haecceitism demands transworld bilocation of individuals; unmysterious quidditism demands transworld bilocation of properties.

But bilocation of individuals, whether between worlds or times or places, is trouble. For bilocated individuals are apt to have different intrinsic properties at their different locations. For instance, A may be standing and B sitting in \(W_1\), whereas in the indiscernible world \(W_2\) it is B who is standing and A who is sitting. How can we make sense of an intrinsic difference between something and itself?16 The problem is serious enough that we should not believe in bilocated individuals at all, in which case haecceitistic differences are utterly mysterious. Bilocated properties raise no similar problem: I can think of no plausible example of an intrinsic (higher-order) property which a bilocated property has at one but not another of its locations.17 Hence it is unproblematic to think of properties as bilocated in time and space. Indeed, we often do think of them that way, whether we believe they are immanent universals wholly present in different locations, or whether we believe they are classes of possible individuals or of resembling tropes with members at different locations. Transworld bilocation of properties is equally unproblematic. We should be quidditists and not haecceitists because haecceitism, unlike quidditism, carries an unaffordable price.
It would be possible to combine my realism about possible worlds with antiquidditism. I could simply insist that, despite the undoubted ability of properties to indulge in bilocation, no property is ever instantiated in two different worlds. That would get in the way of transworld similarity, which we might have thought was transworld sharing of properties. But I could say for world-bound properties, as I do for world-bound individuals, that a counterpart relation is a substitute for transworld identity: two world-bound properties in two worlds are counterparts if the role of one in one world approximately matches the role of the other in the other. Then we could say that individuals in the two worlds are alike to the extent that the properties of one are counterparts of the properties of the other. But although I could treat individuals and properties alike in this way, why should I want to? It could be for the sake of upholding identity of structurally indiscernible worlds, but I see no good reason for wanting to uphold that principle. Or it could be for the sake of blocking the argument for Humility. But why should I want to block that argument? Why is Humility “ominous”? Who ever promised me that I was capable in principle of knowing everything?

The situation changes radically if we are actualists: if we suppose, as many do, that possible worlds are constructions out of actual materials. We might say, for instance, that they are state descriptions in a “Lagadonian” language in which each actual individual and each actually instantiated fundamental property serves as a name for itself (see Lewis 1986a, 145ff.). On the one hand, for an actualist, properties are not really instantiated within worlds; rather a world represents them as being instantiated. So the bilocation of properties within the actual world is now no precedent for transworld bilocation. On the other hand, we have an easy way to account for haecceitistic distinctions involving actual individuals and actually instantiated properties. If worlds mention individuals by (Lagadonian) name, W2 can simply say that B rather than A occupies a certain place. Quidditistic distinctions are just as easy: W2 can simply say that G rather than F occupies a certain place. If it is this easy to capture haecceitistic and quidditistic distinctions, the questions of haecceitism and quidditism can be settled—if they can be settled at all—just by deciding whether we’re moved more by the \textit{prima facie} intelligibility of such distinctions or by our dislike of distinctions without a difference.

But what an actualist cannot do is to capture haecceitistic distinctions involving alien individuals—unactualized individuals which cannot be described as recombinations of actually existing parts; or quidditistic distinctions involving alien fundamental properties. The trouble is that alien
individuals and properties are not available within actuality to serve as Lagadonian names for themselves, nor can we construct any other names for them. An actualist faces a choice: he can reject aliens, or he can reject haecceitism and quidditism. (Or he can reject combinatorialism, without which prima facie examples of haecceitistic and quiditistic distinctions would not arise.)

5 A Second Argument for Humility

Our first argument for Humility—call it the permutation argument—has an advantage and a drawback. The advantage is that it makes use of only the most unquestionable of fundamental properties: those that play an active part in the actual workings of nature. The drawback is that it establishes Humility only for those fundamental properties that fall in more-than-one-membered categories. (In the unlikely event that no two fundamental properties fell in the same category, none would be eligible for permutation, so the argument would not establish any sort of Humility at all.) Now I shall give a second argument that has neither the advantage nor the drawback. Whether that is on the whole an improvement, I shall not say.

The skeleton of the second argument—call it the replacement argument—is the same as before. We start with the unique actual realization of T; all fundamental properties except idlers and aliens are members of it. If we replace those properties by others, we get a possible realization by combinatorialism, and it is different from the one we had before by quidditism. Since no possible observation gives us evidence that goes beyond the Ramsey sentence, we have no way to tell which possible realization is actual.

The difference lies in where we find replacements for the members of the original realization. Instead of permuting those members, we now draw instead on idlers or aliens. Given an abundant supply of idlers or aliens or both, we can replace every fundamental property that is a member of the original realization, whether or not it is the sole member of its category. But should we believe in abundant idlers or aliens?

Begin with aliens. One who believes that ours is one world among many has two reasons to believe in abundant aliens. Both rest on the plausible claim that it is a contingent matter which fundamental properties are instantiated, so it ought to be possible for either more or fewer of them to be instantiated than actually are. So, first, once we appreciate that it is a contingent matter which of them are instantiated, it seems that there
ought to be a world where more of them are instantiated than are instantiated in our world. So properties are instantiated in that world which are alien to ours. And if we take a world where more are instantiated, it seems possible at that world that still more should be instantiated. And so on. What might be questioned is whether there is any world where all fundamental properties are instantiated. Is there room for all of them in any one world? But surely our world is nowhere near running out of room. Second, start with our actual world. Any of the actually instantiated fundamental properties have been uninstantiated. So properties that are instantiated at our world are aliens at other worlds. But if there are worlds where some properties are aliens, and ours is not one of them, then our world has a special distinction that some other worlds lack. Why should the world we happen to live in be special in this way? Very likely it isn’t. I think this second reason is weaker than the first, but still a strong reason for me to believe in aliens.

My view that ours is one world among many is contentious, to say the least. What should you think about aliens if you are an actualist? You do think that our world is special, so my second reason vanishes. My first retains at least some force. However, we saw earlier that an actualist faces a choice between rejecting aliens and rejecting quidditism (or combinatorialism). Perhaps, on balance, rejecting aliens is the wiser choice. Anyway, it doesn’t matter which choice he makes. To run the argument from replacement using aliens, he needs a supply of aliens and he also needs combinatorialism and quidditism. So the argument is not for him.

Next, idlers. Surely the laws of nature might have been different in such a way that some of the fundamental properties which actually play an active part in the workings of nature would have been idlers instead. So idlers are possible. At this point, I can say that if there are no idlers in our world, then our world has a special distinction that some other worlds lack. Why should the one world among many that we happen to live in be special in this way? Very likely it isn’t. This is at best an inconclusive reason for me, since after all it is a contingent question whether our world has idlers and I have no real evidence. If you are an actualist, it is no reason at all for you, since you do think that our world is special. I don’t see that an actualist has any reason at all to believe in idlers.

But how much reason has he to disbelieve in idlers? I grant that Occam’s razor justifies him in having a low degree of belief in idlers. But can he claim to know that there are no idlers? I think not. To say they don’t exist because we can have no reason to believe they do exist seems nothing better than an appeal to verificationism.
The upshot is that whatever you think about the metaphysics of modality, you should at least agree that for all you know, there may be an abundant supply of idlers. So for all you know, the replacement argument using idlers may succeed. *But that is good enough.* Humility is, after all, a thesis of irremediable ignorance. If we are irremediably ignorant about whether the replacement argument using idlers establishes Humility, then either the argument does establish it, and it is true; or else the argument does not establish it, but we cannot know that. In both cases alike, we are irremediably ignorant about the identities of the fundamental properties that figure in the actual realization of the true final theory.

6 Spreading Humility

So far, we have been concerned with Humility about the fundamental properties. But the great majority of properties, including most of the fairly natural properties, most of the properties we are acquainted with and have names for, and most of the intrinsic properties that Langton’s Kant thought were unknown to us, are not fundamental.

I said that the fundamental properties (including relations and magnitudes) are a minimal basis on which all else supervenes. ‘All else’ should at least include all qualitative properties of things, natural or not, intrinsic or not. (Perhaps this should be taken as the definition of ‘qualitative property’.)

First, take the case of a qualitative property instantiated by the entire world. It will supervene on the pattern of instantiation of fundamental properties by the parts of the world. Unless it is itself a fundamental property, it will be either a structural property constructed out of the fundamental properties, or else a truth-functional compound, finite or infinite, of such structural properties.

Next, take the case of an intrinsic property instantiated by some part of the world. Since it is intrinsic, this same property could have been instantiated by something that did not coexist with anything distinct from itself—in other words, by the entire world (see Langton and Lewis 1998). So this case reduces to the previous one. The property is either fundamental, or structural, or a truth-functional compound of structural properties.

Third, take the case of an extrinsic property of something. There will be some more inclusive thing such that the smaller thing has the extrinsic property because the larger thing has some intrinsic property. (In some cases the more inclusive thing will be the entire world, in other cases not.)
For instance, I have the extrinsic property of being an uncle in virtue of an intrinsic property of my family. So this case reduces to the previous one.

Suppose we have a structural property: say, the property of being composed of an F bearing relation R to a G, where F, R, and G are fundamental. If we do not know the identities of the properties we have called ‘F’, ‘R’, and ‘G’ we also do not know the identity of the structural property constructed out of them. If the so-called F might for all we know be either F₁ or F₂, and R might be either R₁ or R₂, and G might be either G₁ or G₂, there are eight different possibilities for what the structural property might be. The same point applies to more complicated structural properties, and to at least some truth-functional compounds of structural properties. So if Humility applies to some or all of the fundamental properties, then it applies also to very many more properties. We may know them as role-occupants, including both their roles in scientific theory and their roles in daily life, and that might pass for “knowing what they are” by lax and commonplace standards. But if they are structural properties and we don’t know the recipe for their construction out of fundamental properties, or if they are certain truth-functional compounds of structural properties and we don’t know the recipes for those structural properties (or we don’t know, even in a lax sense, what structural properties they are compounded from), or if they are unidentified fundamental properties, then in a more demanding sense we do not know what they are, no matter how familiar we may be with them.

There are some exceptional truth-functional compounds, however, to which Humility does not apply. We saw that being composed of an F bearing R to a G might be any of eight different structural properties. We can identify the disjunction of all eight, even if we can’t identify any one of its disjuncts. Likewise, even if we cannot identify the fundamental or structural property that actually occupies a certain role, we can identify the property of having whatever property it is that occupies that role (in the case in question). Perhaps it is thanks to such properties that O-language manages to be fully interpreted.

7 Ineffable Ignorance

Suppose we know of some property only as the occupant of a role, perhaps the role of being the seventeenth member of the actual realization of T, or perhaps some role that is less cumbersome to describe. Suppose Humility applies to this property. We cannot answer the question: which property
occupies that role? But worse: not only can we not answer that question, we can’t even ask it.

How’s that? Didn’t I ask it just now? Yes, but I didn’t ask it properly. To ask it properly is to ask it as a multiple-choice question: is it this, that, or the other? It is to ask it in such a way that, apart from limitations of finitude, the addressee could list all the alternative possible answers to it (see Belnap and Steel 1976). That is what I did not do, and that is what we cannot do.

There are alternative answer-propositions, to be sure. For each fundamental property F (of the right category), there is a contingent proposition true at all and only the worlds where F occupies the role in question. But we do not have alternative answer-sentences that express those alternative answer-propositions and do so in such a way that we can know which sentence expresses which proposition.

“Which property occupies the role?—The occupant of the role, whatever that is.” A true answer, sure enough, but not an answer to the question we meant to be asking. Indeed, not an answer to any question we’d be likely to ask: the only information it conveys is that the role is uniquely occupied.21

Reference-fixing will not help. To make it easy for ourselves, assume we have descriptions to rigidify. Assume we are—rashly—convinced of the nonexistence of both idlers and aliens, so the only fundamental properties that exist are those we can describe as ‘the so-and-so member of the actual realization of T’—for short: ‘the occupant of the so-and-so role’. Assume for simplicity that we have only one category and that all members of the actual realization of T are fundamental properties. Use ‘actual’ as a rigidifier. Now can we ask our multiple-choice question thus?

Which property occupies the seventeenth role? Is it the actual occupant of the first role? Is it the actual occupant of the second role? . . . Is it the actual occupant of the seventeenth role? . . . Is it the actual occupant of the nth role?

No sooner do we ask our question this way than we seem to know the answer: it is the actual occupant of the seventeenth role. But we have cheated. Each of our answer-sentences does indeed express one of the alternative contingent answer-propositions, but we do not know which sentence expresses which proposition. We know that ‘It is the occupant of the seventeenth role’ expresses the true answer-proposition, but that is no help in knowing which one is true. We have fixed the reference blindly, not knowing which rigidified descriptions were rigid designators of which fundamental properties.22
8 Humility about Qualia

A friend of phenomenal qualia might speculate that all the actually instantiated fundamental properties are qualia. That would not preclude them from also occupying physical roles—a sort of panpsychism. So, even if our true final theory is a physical theory, they might be the members of its actual realization. He might also accept the Identification Thesis: anyone acquainted with a quale knows just which property it is. Now it may seem that we can know the identities of the fundamental properties after all—we need only become acquainted with them.23

Since he rejects the conclusion of our arguments for Humility, the friend of qualia had better reject some assumption that went into those arguments. He has two choices for what to reject. (1) I assumed that O-language names fundamental properties only as occupants of roles, and that T tells us that the referent of so-and-so T-term occupies such-and-such role. The friend of qualia may say that all fundamental properties are named in O-language, and not as occupants of roles, provided we take O-language to be the inner private language of someone who has become acquainted with certain qualia. Or (2) if he would prefer to have nothing to do with the dubious notion of an inner private language, he could instead take issue with our assumption that O-language suffices to express all possible observations. He could say that it does not suffice to express those observations that consist in gaining acquaintance with qualia.

As a materialist, I reply predictably: the Identification Thesis is false (see Lewis 1995). If the Identification Thesis is built into the very definition of qualia, there are no qualia. If not, there may well be qualia, but they are known to us only as role-occupants, and there are multiple possible occupants of the roles in question. Their roles are psychological: for instance, qualia confer on us abilities to recognize and imagine what we have previously experienced. But we cannot tell just by acquaintance which property occupies such a role. Obviously not, if the occupant of the role is some complicated neural property. But not if it is a fundamental property either. In that case, Humility applies to qualia just as it would to any other fundamental properties that were known to us only as role-occupants.

Even if the Identification Thesis is false for us, might it be true for some other being, say God? Indeed, might God have the supernatural power to become acquainted with all fundamental properties, qualia or not, and to identify each of them by acquaintance? And if Humility did not apply to God, would He then be able to tell us just which fundamental property it is that occupies any given role?
If there were a God, who knows what supernatural powers He might have. But no matter what knowledge He might gain by acquaintance with the fundamental properties, He could not share it with us. Since we cannot express any of the answer-propositions to the question which fundamental property occupies a given role—not in such a way that we know which of the propositions we are expressing—God would have no way to communicate His knowledge to us. If He wanted to remedy our ignorance, His only recourse would be to impart to us His own power to identify properties by acquaintance.

Acknowledgments

Thanks are due to Karen Bennett, Robert Black, Cian Dorr, Allan Gibbard, Charles Goodman, John Hawthorne, Anja Jauernig, Rae Langton, Jonathan Schaffer, Ted Sider, and Dean Zimmerman.

Notes

1. See, inter alia, Jackson, Pargetter, and Prior 1982; Prior 1985; Armstrong 1997, 69ff.; and Lewis 1997 (reprinted as chapter 7 in Lewis 1999). These authors disagree about whether a disposition should be identified in each particular case with its categorical ground in that case, or whether instead it is the disjunctive property of having some categorical ground or other. But that is a merely terminological question and need not concern us here.

2. See Lewis 1983a. Neither name has proved fully satisfactory. ‘Fundamental properties’ suggests to some that they must be properties of fundamental particles. They might be; or they might instead be properties of space-time points, or of extended things, or even of unlocated things. ‘Natural properties’ suggests to some that it is nature that makes them natural: in other words, that it is contingent which properties are natural. That was not my intention. The properties that figure in the fundamental laws of nature are natural, but that is not because figuring in the laws makes them natural. Rather it is because regularities are fit to compete for the status of lawhood only when formulated in terms of perfectly natural properties (ibid., 367ff.).

3. That creates a problem about the properties instantiated in an infinitely complex world of “structures all the way down.” I take infinite complexity to be a genuine possibility, though a far-fetched one. (I don’t know what to say to someone who thinks it is not far-fetched.) For it may be that all of the properties instantiated in such a world are structural. If so, and if fundamental properties are never structural, we must conclude that in such a world no perfectly fundamental properties are instantiated, but only near-enough fundamental properties. We could then
conclude that some of the privileges otherwise reserved for perfectly fundamental properties can belong instead to the near-enough fundamental properties that are instantiated in infinitely complex worlds: for instance, the privilege of appearing in fundamental laws of nature, or the privilege of corresponding to universals or tropes, or the privilege of constituting a basis—not, however, a minimal basis—on which all else supervenes. I see no reason to object to these conclusions.

4. In my view, this is the proper conclusion to draw about the famous Quinean circle from analyticity to synonymy and back. Quine is right that there is a problem about analyticity, but the problem is not the circle. Rather it is that in most interesting cases, candidates for analyticity are analytic under some legitimate resolutions of semantic indeterminacy and not under others.

5. Optimists hope and expect that we will discover the final theory someday soon, or anyway someday. I share the hope and expectation, but I am not assuming it. Maybe scientific research will go out of fashion. Maybe the task of fully understanding the workings of nature is just too hard for us. But whether or not we will ever discover the final theory, it nevertheless exists at least in the way never-to-be-written poems do. Materialists hope and expect that the true final theory, whether or not we ever discover it, will be a physical theory: it will take over large parts of current physics, perhaps in corrected form, and it will be in the distinctive style of current physics. For instance, it will explain the workings of large things in terms of the workings of their very small parts; it will get by with very few fundamental properties and laws; and it will not posit different fundamental properties and laws for the conscious and the unconscious parts of nature, or for the living and the dead parts of nature. Again I share the hope and expectation, but again I am not assuming it.

6. Proof. Let \( S \) be any \( O \)-language sentence. If the Ramsey sentence implies \( S \), so does the postulate, because the postulate implies the Ramsey sentence by Existential Generalization. Conversely, suppose the postulate implies \( S \).

\[
\begin{align*}
(1) & \quad \text{Then } [T(t_1, \ldots, t_n) \text{ only if } S] \text{ is a logical truth.} \\
(2) & \quad \text{Then All } x_1, \ldots, \text{ All } x_n [T(x_1, \ldots, x_n) \text{ only if } S] \text{ is a logical truth.} \\
(3) & \quad \text{Then (Some } x_1, \ldots, \text{ Some } x_n [T(x_1, \ldots, x_n)] \text{ only if } S) \text{ is a logical truth.} \\
(4) & \quad \text{Then the Ramsey sentence implies } S.
\end{align*}
\]

The only step that requires comment is the step from (1) to (2). This step is an instance of the rule of Universal Generalization—a rule which, though it does not preserve truth, does preserve logical truth in any system of logic worthy of the name. Its rationale is that logic does not discriminate between things. The only way to have a logical truth, \(-t-\), about a particular named thing (when the name ‘t’ is neither structured nor an item of logical vocabulary, as indeed our T-terms are not) is to have it as an instance of a general logical truth: \( \forall x [-x-] \). Note that neither this nor any other step of the proof depends on any idiosyncrasies of first-order logic, or extensional logic, or finitary logic.
7. *Proof.* The Ramsey and Carnap sentences jointly imply the postulate by *Modus Ponens*. The postulate implies the Ramsey sentence by Existential Generalization, as already noted. The postulate implies the Carnap sentence because it is the consequent of the Carnap sentence.

8. *Proof.* We proved that the Ramsey sentence of $T(t_1, \ldots, t_n)$ implies the same O-sentences as $T(t_1, \ldots, t_n)$ itself. $T(t_1, \ldots, t_n)$ needn’t be the postulate of $T$; the proof works no matter what sentence it may be. So in particular the Ramsey sentence of the Carnap sentence implies the same O-sentences as the Carnap sentence itself. But the Ramsey sentence of the Carnap sentence is a logical truth, so it implies no O-sentences except logical truths. So the same is true of the Carnap sentence. Again, the proof does not depend on any idiosyncrasies of first-order logic, extensional logic, or finitary logic.

9. I once proposed adding that if a theory has no realization, or multiple realizations, its theoretical terms do not refer (Lewis 1970). I’d now say that if it is unrealized but almost realized, its theoretical terms refer to the members of its unique near-realization, if there is one; and that if it has multiple realizations (or near-realizations) its theoretical terms have indeterminate reference. My reason for saying that the theoretical terms of a multiply realized theory do not refer was that a theorist may be presumed to have intended to implicitly define the terms he introduces. But there is a simpler way to respect the theorist’s presumed intention: we should write the postulate in such a way that his theory cannot be multiply realized.


11. If laws of nature are suitable regularities, the exchange of locations of $P_1$ and $P_2$ will guarantee the exchange of nomological roles. If laws of nature are certain special law-making relations of fundamental properties, $P_1$ and $P_2$ will have to be exchanged also as arguments of these law-making relations to guarantee the exchange of nomological roles. On these two conceptions of lawhood, see further endnote 12.

12. One way for the laws of nature to be contingent is for them to be those regularities that earn a place in the true system that achieves the best possible balance of simplicity and strength. See Mill 1843, Book III, chapter IV, section 1; Ramsey 1990, chapter 7A, “Universals of Law and of Fact,” 140–144; and Lewis 1973, 73ff. Another way for laws to be contingent is for there to be a law-making relation $N$ between fundamental properties, such that it is contingent which properties stand in this relation but necessary that whenever $N(F, G)$, every $F$ is a $G$. (Or perhaps what’s necessitated is something a little more complicated that allows laws to be defeasible.) See Dretske 1977; Tooley 1977; Armstrong 1983. Although the first-order law that every $F$ is a $G$ is contingent, the higher-order law that whenever $N(F, G)$, every $F$ is a $G$ is necessary. Whether that violates combinatorialism is a disputed matter.

13. Shoemaker 1980; Swoyer 1982; Ellis and Lierse 1994. Necessity of the laws would imply that combinatorialism was false. It would block our present argument for
Humility, but would not imply that it was false. Contingency of the laws implies nothing one way or the other about the truth of combinatorialism, and nothing one way or the other about Humility.

14. Further, principles of identity of indiscernibles lead to the following question. Nobody doubts that there could be two nonidentical almost-indiscernibles. If they had been just a little bit different, they would have been nonidentical indiscernibles. What stops them from being that little bit different? See Adams 1979.

15. Some seeming cases of haecceitistic difference are compelling enough that an antihaecceitist needs to explain them away. I do so by arguing that not all possibilities are possible worlds; rather, some are possibilities for individuals (or pairs, triples, . . . of individuals); and it is not haecceitism if individual possibilities within a single world differ by a permutation of individuals. See Lewis 1986a, 230ff.

16. I take it that explanations which replace intrinsic monadic properties by relations, or which take individuals not to have their properties simpliciter but only to stand in some relation to them, are unacceptiable. See Lewis 1986a, 199ff., and Lewis 2002.

17. Armstrong (1989, 67) considers and rightly rejects a somewhat plausible case of intrinsic difference between a property in different locations: a universal that is allegedly simple at one world, complex at another.


Plantinga (1974) hopes to side-step the problems of haecceitism about alien individuals and quidditism about alien properties by invoking allegedly actual surrogates for these aliens. Fine (1985) argues convincingly that this will not do because the surrogates presuppose the possibilia they are meant to be surrogates for.


20. By lax and commonplace standards, coming to know what something is means, roughly, acquiring a fuller description of it. You could come to know who Great Leader is by learning that he is the man on yonder balcony giving the clenched-fist salute. Equally, you could come to know who the man giving the salute is by learning that he is Great Leader. Discovering his haecceity, or even the necessary and sufficient conditions for something to be his counterpart, doesn’t enter into it.

21. If that. The role might be multiply occupied, the improper description ‘the occupant of the role’ might be indeterminate in reference, and we might be
following the supervaluationist policy of coping with semantic indeterminacy by saying what’s true on all resolutions of the indeterminacy.

22. The real lesson of Kripke’s alleged examples of *a posteriori* necessity and *a priori* contingency (Kripke 1980, 97ff., 128ff.), is that when we have reference fixing, we have two different ways for a sentence to express a proposition. It may happen that the proposition expressed one way is necessary and knowable *a priori* but the proposition expressed the other way is contingent and knowable only *a posteriori*.

See Chalmers 1996, 56ff.; Stalnaker 1978; Jackson 1998a, 46ff. Let sentence S contain terms subject to reference fixing. (Ignore other forms of indexicality.) The *primary intension* of S (Chalmers), or the *diagonal proposition* (Stalnaker), or the *A-intension* (Jackson) is the proposition that is true at just those worlds W such that S is true at W with respect to reference fixing as it is at W. The *secondary intension* of S (Chalmers), or the *horizontal proposition* (Stalnaker), or the *C-intension* (Jackson) is the proposition that is true at just those worlds W such that S is true at W with respect to reference fixing as it is in actuality. If, for instance, the reference of ‘heat’ is fixed on molecular motion at the actual world but on other things at other worlds, the primary intension of ‘Heat is molecular motion’ is contingent but the secondary intension is necessary; and *vice versa* for ‘Heat is the occupant of the heat-role’.

So far, when I have spoken of the propositions expressed by the answer-sentences, I’ve meant the secondary intensions. If instead we took the primary intensions, it would be just as if we’d omitted the rigidifier ‘actual’. As already noted, that would not be a way to ask the question we meant to ask. We would know the primary intensions of the answer-sentences, but these would not be the answer-propositions that identify the properties that occupy the roles.

23. See Maxwell 1978; and for a sympathetic discussion of the view, see Chalmers 1996, 135ff. Just as our panpsychist thinks it is because fundamental properties are qualia that we can know their identities by acquaintance, so others think it is because (or insofar as) they are not qualia that we cannot know their identities. See, *inter alia*, Russell 1927, 497; Unger 1999.
A Partial Defense of Ramseyan Humility

Dustin Locke

1 Introduction

In “Ramseyan Humility” (this volume), David Lewis argues that “we are irremediably ignorant about the identities of the fundamental properties that figure in the actual realization of the true final theory” (214). Of the three published responses to Lewis’s essay (Langton 2004; Schaffer 2005; Whittle 2006), each argues that even if we accept Lewis’s metaphysical assumption (the thesis known as ‘quidditism’), we need not accept his epistemic conclusion (hereafter, ‘Humility’). My aim in this paper is to defend Lewis against these critics.

Ann Whittle claims that Humility rests on an implausibly strong account of identification—namely, that to identify X, one must be able to distinguish X from all other actual and possible entities. Accordingly, Whittle attempts to refute Humility by an appeal to a more lenient account of identification. In what follows I defend Lewis by showing that Lewis’s demanding account of identification is a perfectly good account of at least one perfectly legitimate sense of ‘identification’.

Jonathan Schaffer and Rae Langton each claim that Humility rests on an implausibly strong account of knowledge—namely, noncontextual infallibilism. Accordingly, Schaffer and Langton attempt to refute Humility by appealing to more lenient epistemic principles. In what follows I defend Lewis, not by defending noncontextual infallibilism, but by showing that even on the alternative epistemic principles of Schaffer and Langton, Lewis’s conclusion still follows.

2 The Argument for Humility

Lewis’s argument can be conveniently reconstructed in five stages.
The stage-setting stage. Let T be a true and complete final theory of our world. Lewis claims that the language of T can be divided into two parts: the terms that have meaning independently of T (the ‘O-terms’) and the terms that are implicitly defined by T (the ‘T-terms’). Writing T as a single sentence and replacing the T-terms with variables, we get the realization formula of T—the open sentence ‘T(x₁, x₂, . . . , xₙ)’. Any n-tuple of properties that could have satisfied the realization formula of T is called a ‘possible realization of T’; any n-tuple of properties that does satisfy the realization formula of T is called an ‘actual realization of T’.

The metaphysical/linguistic stage. Lewis assumes that there is a realization of T, say, ⟨P₁, P₂, . . . Pₙ⟩ such that at least two of P₁, P₂, . . . Pₙ are fundamental properties of the same adicy (1-place, 2-place, etc.) and category (all-or-nothing properties, real-valued magnitudes, etc.). He further assumes that if the O-language refers to the properties that realize T, it is only via descriptions of their nomological/locational roles in nature. Finally, Lewis assumes quidditism—the thesis that for any fundamental property P and nomological/locational role of appropriate adicy and category R, P might have realized R. Given these assumptions, it follows that there is some permutation of the actual realization of T that is a possible realization of T. Since the actual realization of T is of course also a possible realization of T, it follows that T has multiple possible realizations.

The first epistemic stage. Lewis assumes that the O-language of T suffices to describe all possible observations. Now consider the Ramsey sentence of T—‘∃x₁, x₂, . . . , xₙT(x₁, x₂, . . . , xₙ)’. Since the Ramsey sentence of T implies all and only the O-language sentences of T, it follows from the above assumption that our (observational) evidence will always be consistent with the proposition expressed by the Ramsey sentence of T. Thus, for any world w such that the proposition expressed by the Ramsey sentence of T is true at w, w will never be eliminated by—that is, inconsistent with—our evidence.

The second epistemic stage. Although Lewis doesn’t say so, the most straightforward way to get from the conclusion of the previous stage to the conclusion of Ramseyan Humility is via Lewis’s own account of knowledge. In Lewis 1999, Lewis defends the following infallibilist account of knowledge: S knows that P iff S’s evidence eliminates all possibilities in which not-P (psst!—see this note). This analysis, together with the conclusion of the last stage of the argument, implies that we will never know p, where p is the proposition that is true at all and only those worlds where the actual realization of T is the realization of T (= the proposition that is true at all and only those worlds where ⟨P₁, P₂, . . . Pₙ⟩ realizes T).
The third epistemic stage. The final stage of the argument takes us from the claim that we will never know \( p \) (from above) to the claim that we will never know which \( n \)-tuple of properties realizes \( T \). As in the last stage of the argument, Lewis leaves the key assumption here implicit, but the path seems fairly clear. We first assume what Jonathan Schaffer (2007) has called the “Standard Analysis of knowing-\( wh \),” according to which \( S \) knows-\( wh \) (who, which, what, etc.) iff \( S \) knows \( P \), where \( P \) is the true answer to the indirect question of the \( wh \)-clause. If we assume also that \( p \) is the true answer to the question ‘Which \( n \)-tuple of properties realizes \( T \)?’ it follows from the conclusion of the previous stage of Lewis’s argument that we will never know which \( n \)-tuple of properties realizes \( T \).

As I said above, Whittle, Schaffer, and Langton all argue that even if Lewis’s metaphysical assumption (quidditism) is correct, his conclusion (Humility) can be blocked. We have just seen that Lewis’s conclusion cannot be blocked if we assume the noncontextual version of Lewis’s infallibilist account of knowledge and the account of knowing-\( wh \) given in the third epistemic stage. But Whittle, Schaffer, and Langton are by no means obligated to grant Lewis these accounts. After all, the account of knowledge is an infallibilist account of knowledge and the account of knowing-\( wh \) is a quite strong one. Accordingly, we might have expected these accounts to imply some pretty serious epistemic depravity. The interesting question, then, is whether quidditism implies Humility on some more lenient epistemic principles. This is exactly the question that these authors attempt to address, and they all answer it in the negative. We will look at their arguments, and why I think they’re wrong, in a moment. For now, we must pause to clear something up.

3 Interlude: Trivial Knowledge and Two Dimensions of Intension

In the last section I argued that Lewis’s thesis that we will never know which properties realize \( T \) is equivalent to the thesis that we will never know \( p \), where \( p \) is the proposition that is true at all and only those worlds where the actual realization of \( T \) is the realization of \( T \). Now this may seem a bit odd, for this is just the thesis that

(1) We will never know that the actual realization of \( T \) is the realization of \( T \).

But how could (1) possibly be true? What could be easier than knowing that the actual \( F \) is the \( F \)?
This response hinges on an ambiguity in (1). The ambiguity concerns which proposition is expressed by the embedded sentence ‘the actual realization of T is the realization of T’. One way to understand the ambiguity is with the resources of “nonambitious” two-dimensional semantics. There are several particular versions of two-dimensional semantics, but only the general idea is needed here. This is the idea that at least some statements have two intensions: a C-intension and an A-intension. The A-intension of a statement, roughly, is the proposition that is true at all and only those worlds considered as actual where the statement is true (‘A’ for ‘actual’). The C-intension, on the other hand, is the proposition that is true at all and only those worlds considered as counterfactual where the statement is true (‘C’ for ‘counterfactual’).

Let’s take an example. Consider the statement ‘The actual US president is the US president’. Now consider a world w1 where the US president is Al Gore. Is the statement ‘The actual US president is the US president’ true at w1? Well, that depends on whether we’re considering w1 as actual or as counterfactual. Consider it first as counterfactual. In other words, ask yourself what would have been true had Al Gore been the US president. In particular, ask yourself if the actual US president would have been the US president. The answer, it seems, is ‘no’—the actual president, George W. Bush, would not have been the president had Al Gore been the president. But now consider w1 as actual. In other words, ask yourself what is true if Al Gore is the US president. In particular, ask yourself if the actual US president is the US president if Al Gore is the US president. The answer in this case, it seems, is ‘yes’. Thus, the A-intension and C-intension of ‘The actual US president is the US president’ are not identical: the former is true at w1 but the latter is not.

Exactly the same thing happens in the case of ‘The actual realization of T is the realization of T’. The A-intension of this statement is true at all and only those worlds where T is uniquely realized, regardless of which properties realize it at that world. This is simply because for any world w where T is uniquely realized, the statement ‘The actual realization of T is the realization of T’ is true at w considered as actual. However, the C-intension of ‘The actual realization of T is the realization of T’ is not true at every world where T is uniquely realized. In particular, it is false at any world where T does not have the same realization that it does in this (the actual) world. This allows us to explain the ambiguity in (1). If the embedded statement in (1) expresses its A-intension, then (1) is false, provided only that we will come to know that T is uniquely realized. However, only if the embedded statement in (1) expresses its C-intension is (1) equivalent to Lewis’s thesis.
This concludes my explication of Lewis’s thesis and his argument for it. I turn now to the criticisms of Lewis and my defense.

4 Whittle’s Response to Humility

Ann Whittle argues that even if Lewis’s metaphysical assumptions are correct, “we can [still] know which property occupies a certain role, because we are able to *identify* the property in question” (2006, 469, emphasis added). I disagree.

Whittle’s argument explicitly assumes a version of Russell’s principle: being able to identify A is having ‘discriminating knowledge’ of A, where one has discriminating knowledge of A if one has a description that uniquely picks out A. Thus, since we can uniquely pick out the property that realizes such-and-such role with the description ‘the property that realizes such-and-such role’, it follows by Whittle’s version of Russell’s principle that we can identify the property that realizes such and such role.

The first thing to notice about Whittle’s version of Russell’s principle is just how cheap it makes identification. An example should make this clear. Suppose we’re in a courtroom and I’m on the stand. An attorney asks me, “Do you know which person stole your chicken?” Suppose that all I know about the person who stole my chicken is that the person who stole my chicken is the person who stole my chicken. However, it turns out that there is a unique person who stole my chicken. It follows by Whittle’s version of Russell’s principle that I can identify the person who stole my chicken.

Clearly Lewis had a stronger notion of ‘identification’ in mind when he stated his thesis. Whittle seems to agree. She writes, “Lewis, however, seems to think that something more is required: that we be able to single out the entity from all other actual and possible entities” (ibid., 470). But she goes on to argue that

[S]uch a condition on identification seems too strong. Suppose, for instance, that I am a historian who knows all there is to know about Napoleon and this knowledge, we can safely assume, allows me to single out one person in history, namely Napoleon. Granted a certain view about the transworld identities of particulars, namely Haecceitism, this exhaustive knowledge of Napoleon would, nevertheless, fail to identify him. Why? Haecceitism states that two possibilities can differ just in the permutation of individuals. So everything could be qualitatively identical in two possible worlds, yet those worlds differ in that the person we name ‘Napoleon’ instantiates all the same properties that Nelson does in this world, and vice versa. (Ibid.)
Whittle’s mistake, it seems to me, is in searching for one (or perhaps a few) notions that deserve the title ‘identification’. As many theorists have noted, expressions such as ‘identification’, ‘knowing what’, and ‘knowing which’ are all highly context-sensitive. Lewis himself notes that in addition to all the common and less demanding senses of these expressions, there is “an uncommonly demanding and literal sense” (Lewis 1995, 142). Moreover, Lewis explicitly notes (this volume, 215 and endnote 20) that it is this “uncommonly demanding sense” that he has in mind when he says that we will never know which properties realize T.

Above I stated that according to the standard analysis of knowing-\textit{wh}, \( S \) knows-\textit{wh} iff \( S \) knows \( P \), where \( P \) is the true answer to the indirect question \( Q \) of the \( \text{wh} \)-clause. The different senses of ‘knows-\textit{wh}’ can be modeled using this analysis by varying what counts as the true answer to \( Q \). If we want a more demanding sense, we simply require that only more specific propositions count as answers to \( Q \). Taking this process to its limit, we get the “uncommonly demanding and literal sense” of knowing-\textit{wh}. In this sense of knowing-\textit{wh}, to know-\textit{wh} is to know the maximally specific answer to the question. In our case, the question is ‘Which possible realization of \( T \) is the realization of \( T \)?’ and the maximally specific answer is the proposition that \( \langle P_1, P_2, \ldots, P_n \rangle \) realizes \( T \), which, by the second epistemic stage of Lewis’s argument, is a proposition that we will never know.

In any case, whether or not Lewis is correct to state his thesis in terms of ‘knowing which’ or ‘identification’—that is, whether the third epistemic stage of Lewis’s argument is sound—the more interesting question would seem to be whether the argument through the second epistemic stage is sound. After all, the second epistemic stage itself concludes with a claim about our irremediable ignorance. Let us then move on to a discussion of Schaffer (2005) and Langton (2004), whose criticisms appear to be directed precisely at this earlier stage of the argument.

5 This Ain’t Your Daddy’s Skepticism

In “Quiddistic Knowledge,” Jonathan Schaffer argues that Lewis’s thesis “is just a species of skepticism about the external world” and that, therefore, “whatever answer one offers to skepticism about the external world will thereby answer quiddistic skepticism” (Schaffer 2005, 19).

I believe that Schaffer is wrong that Humility is “just a species of skepticism about the external world.” Moreover, I believe that the way in which Humility differs from traditional external-world skepticism is precisely
what makes Humility *immune* to almost all of the typical responses to traditional skepticism. Let me first illustrate the difference between Humility and traditional skepticism by discussing two of Schaffer’s attempts to apply a standard response to traditional skepticism to Humility. After I discuss these particular responses, I’ll attempt to say in somewhat more precise terms what exactly the difference is between Humility and traditional skepticism and why most standard responses to the latter will therefore be unsuccessful responses to the former.

### 5.1 Deductionism (aka the Moorean Response)

Consider what Schaffer calls the “deductionist” response to external-world skepticism. A deductionist response takes the following form:

1. I know that \( p \) (by common sense).
2. If I know that \( p \), then I am in a position to know that \( q \).
3. Therefore, I am in a position to know that \( q \).

where \( p \) is such that it is “common sense” that I know that \( p \), and \( p \) entails \( q \). If we add to the above argument the assumption that I have deduced \( q \) from \( p \), it follows that I know that \( q \).

For example, the deductionist might respond to the brain-in-a-vat skeptic as follows:

1. I know that I have hands (by common sense).
2. If I know that I have hands, then I am in a position to know that I am not a handless brain in a vat.
3. Therefore, I am in a position to know that I am not a handless brain in a vat.

Might we offer a deductionist response to Humility? Schaffer thinks so: “[Just] as the deductionist claims that by starting with one’s knowledge that one has hands, one can come to know that the external world is real, so she should claim that by starting with one’s knowledge that, for instance, this brick has mass, one can come to know the quiddities. The same deductive moves are available in both cases” (2005, 21). So Schaffer thinks that the deductionist should offer something like the following reply to Humility:

1. I know that this brick has mass (by common sense).
2. If I know that this brick has mass, then I am in a position to know that mass realizes the mass role.
3. Therefore, I am in a position to know that mass realizes the mass role.

Will this line of argument work? If so, it must be “common sense” that I know that this brick has mass. Suppose, on the deductionist’s behalf, that it is common sense that I know that this brick has mass. What proposition do I thus know?

I stated earlier that T implicitly defines the T-terms. According to Lewis (1970, 1972, and this volume) a T-term $t_i$ is equivalent to the definite description ‘the $x_i$ such that there exist unique $x_1, x_2, \ldots, x_n$ such that $T(x_1, x_2, \ldots, x_i, \ldots, x_n)$’—for short, ‘the property that realizes the $t_i$ role’. Such definitions are commonly known as ‘Ramsey-style definitions’. Here is one variation of the story that Lewis often told to motivate the idea that T-terms have Ramsey-style definitions:

We are assembled in the drawing room of the country house; the detective reconstructs the crime. That is, he proposes a theory designed to be the best explanation of the phenomena we have observed: the death of Mr. Body, the blood on the wallpaper, the silence of the dog in the night, the clock seventeen minutes fast, and so on. He launches into his story:

$X$, $Y$ and $Z$ conspired to murder Mr. Body. Seventeen years ago, in the gold fields of Uganda, $X$ was Body’s partner. . . . Last week, $Y$ and $Z$ conferred in a bar in Reading. . . . Tuesday night at 11:17, $Y$ went to the attic and set a time bomb . . . Seventeen minutes later, $X$ met $Z$ in the billiard room and gave him the lead pipe . . . Just when the bomb went off in the attic, $X$ fired three shots into the study through the French windows. . . .

And so it goes: a long story. Let us pretend that it is a long conjunctive sentence.

The story contains three names, ‘$X$’, ‘$Y$’ and ‘$Z$’. The detective uses the new terms without explanation, as though we knew what they meant. But we do not. We never used them before, at least not in the senses they bear in the present context. All we know about their meanings is what we gradually gather from the story itself. Call these . . . T-terms . . . because they are introduced by a theory. Call the rest of the terms in the story O-terms. . . .

In telling his story, the detective set forth three roles and said that they were occupied by $X$, $Y$ and $Z$. He must have specified the meanings of the three T-terms ‘$X$’, ‘$Y$’ and ‘$Z$’ thereby; for they had meanings afterwards, they had none before, and nothing else was done to give them meanings. They were introduced by an implicit functional definition, being reserved to name the occupants of the three roles. . . .

If, as I claim, the T-terms are definable as naming the first, second and third occupants of the unique triple that realizes the story, then the T-terms can be treated like definite descriptions. . . . (Lewis 1972, 249–252)

And just in case you were thinking that the descriptions merely fix the referents of the T-terms, without actually giving their semantic contents, Lewis goes on to say
If the story is uniquely realized, they name what they ought to name; if the story is unrealized or multiply realized, they are like improper descriptions. If too many triples realize the story, ‘X’ is like ‘the moon of Mars’; if too few triples—none—realize the story, ‘X’ is like ‘the moon of Venus’. Improper descriptions are not meaningless. Hillary Putnam has objected that on this sort of account of theoretical terms, the theoretical terms of a falsified theory come out meaningless [Putnam 1962]. But they do not, if theoretical terms of unrealized theories are like improper descriptions. ‘The moon of Mars’ and ‘The moon of Venus’ do not (in any way) name anything here in our actual world; but they are not meaningless, because we know very well what they name in certain alternative possible worlds. (Ibid., 252–253)

If Lewis’s semantics of T-terms is correct, and if ‘mass’ is a T-term as we are supposing, then the term ‘mass’ is equivalent to the definite description ‘the property that realizes the mass role’. Let us then reconstruct the deductionist’s argument ((8), (9), and (10)) in light of this equivalence.

(8*) I know that this brick has the property that realizes the mass role (by common sense).
(9*) If I know that this brick has the property that realizes the mass role, then I am in a position to know that the property that realizes the mass role is the property that realizes the mass role.
(10*) Therefore, I am in a position to know that the property that realizes the mass role is the property that realizes the mass role.

According to (10*), what I am in a position to know is that some unique property realizes the mass role. But Lewis’s thesis is that we will never know which property (in the demanding sense) realizes the mass role—that is, we will never know the proposition that is true at all and only those worlds where the property that actually realizes the mass role, is the property that realizes the mass role.

What about rigidifying our descriptions? Consider

(8**) I know that this brick has the property that actually realizes the mass role (by common sense).
(9**) If I know that this brick has the property that actually realizes the mass role, then I am in a position to know that the property that actually realizes the mass role is the property that realizes the mass role.
(10**) Therefore, I am in a position to know that the property that actually realizes the mass role is the property that realizes the mass role.

The conclusion of this argument (10**) indeed contradicts Lewis’s thesis, but only if we read the sentence containing the definite description ‘the
property that actually realizes the mass role’ as expressing its C-intensions (see section 3 above). Is this a move the deductionist can make? No, because in that case the plausibility that the first premise is “common sense” has gone right out the window. What is common sense (if anything here is) is that we know that this brick has mass, which, by Lewis’s semantics for T-terms, is knowing that this brick has the property that realizes the mass role.

5.2 Abductionism (aka Inference to the Best Explanation)

The abductionist response to external-world skepticism claims that (1) the hypothesis that the actual world is roughly as we perceive it to be is a better explanation of our experiences than the hypothesis that, say, we are disembodied dreamers and (2) if the hypothesis that $p$ is the best explanation of one’s experiences, then one can come to know that $p$ by an inference to the best explanation. Schaffer claims that the abductionist ought to have a parallel response to Humility: “[Just] as the abductionist claims that skeptical scenarios constitute poor explanations of the appearances, so she should claim that quiddity swapping scenarios constitute more complex, less conservative, or at least somehow inferior explanations of the powers” (Schaffer 2005, 22). In a footnote to this passage, Schaffer admits that the abductionist will have a difficult time making the case that the “quiddity swapping” scenarios (i.e., the alternative realization scenarios) constitute inferior explanations of the powers. However, he seems to be suggesting that since these same difficulties arise for the abductionist response to traditional skepticism, there is no special difficulty for the abductionist response to Humility.

I disagree. Although there are indeed difficulties for the abductionist’s claim that the normal external-world hypothesis is a better explanation of our experiences than the skeptical hypothesis, at least in that case one can distinguish various features of the competing hypotheses that then figure in the arguments for which is the better explanation. For example, the external-world hypothesis postulates the existence of many things outside of our minds, whereas the dreaming hypothesis does not. But are there any such distinguishing features of the competing hypotheses in the case of Humility?

Let $R_1$ and $R_2$ be two possible realizations of $T$, such that the first and second members of $R_1$ are permuted in $R_2$. Now consider $H_1$, the hypothesis that $R_1$ realizes $T$ and $H_2$, the hypothesis that $R_2$ realizes $T$. What are the distinguishing features of $H_1$ and $H_2$? Well, simply the fact that according to $H_1$, $R_1$ realizes $T$ whereas according to $H_2$, $R_2$ realizes $T$. Is this difference between $H_1$ and $H_2$ relevant to which is the better explanation? There
seems to be absolutely no way to tell, and the reason for this is a fact about how we are forced to represent these two hypotheses.

If \( R_1 \) is the actual realization of \( T \), it may seem that we can represent \( H_1 \) with the sentence ‘\( \langle t_1, t_2, \ldots, t_n \rangle \) realizes \( T(x_1, x_2, \ldots, x_n) \)’, where this sentence uses each \( T \)-term ‘\( t_i \)’. But if we learned our lesson from above, then we’ll know that that won’t do the trick. According to Lewis, the sentence just mentioned is equivalent to the sentence ‘the \( n \)-tuple that realizes \( T \) is the \( n \)-tuple that realizes \( T \)’. This sentence of course is trivial (save its existential and uniqueness presuppositions) and obviously does not express the hypothesis \( H_1 \). What we’ll have to do, if we want to use some sentence like the above to represent \( H_1 \), is rigidify our descriptions. Thus, the sentence ‘the \( n \)-tuple that \( \text{actually} \) realizes \( T \) is the \( n \)-tuple that realizes \( T \)’ manages to express the hypothesis \( H_1 \). Similarly, the sentence ‘the \( n \)-tuple formed by permuting the first and second members of the \( n \)-tuple that \( \text{actually} \) realizes \( T \) realizes \( T \)’ manages to express the hypothesis \( H_2 \).

Our question, remember, is whether \( H_1 \) and \( H_2 \) have any distinguishing features such that we can infer to \( H_1 \) on the grounds that it is a better explanation than \( H_2 \). But now seeing how we are forced to represent these hypotheses, it should be fairly clear that if there are features of \( H_1 \) and \( H_2 \) that make one a better explanation than the other, we are going to be completely ignorant of any such features. As Lewis would say, our rigidified descriptions enable us to represent the hypotheses in question but only in such a way that we are “blinded” to their distinguishing features.

Consider the detective in Lewis’s story above. Suppose that Jack actually realizes the \( X \) role and that Bridget actually realizes the \( Y \) role. Now suppose that the \( \text{only} \) means that the detective has of representing the hypothesis that Jack realizes the \( X \) role and Bridget realizes the \( Y \) role is with the sentence ‘The person who actually realizes the \( X \) role realizes the \( X \) role and the person who actually realizes the \( Y \) role realizes the \( Y \) role’. Can the detective thus infer that proposition (over the proposition that Mary realizes the \( X \) role and John the \( Y \) role) by an inference to the best explanation? On pain of allowing too much knowledge too easily, we had better say “no.” Through his ineliminable use of rigidified descriptions, the detective is blinded to any features of that hypothesis that make it a better explanation than the alternative. The same goes for us in the case of Humility.

5.3 The Diagnosis

At this point we should be asking ourselves questions along the following lines. Don’t we, or might we someday, have a language rich enough to express \( H_1 \) without the use of rigidification? If so, couldn’t we then infer to
that proposition via inference to the best explanation? Moreover, if we have such a language, then perhaps it contains some sentence $S$, such that it is (will be) common sense that we know $S$ (here $S$ is used) and such that $S$ expresses the proposition that this brick has the property that actually realizes the mass role. If so, couldn’t we then deduce the proposition that Lewis claims we don’t know?

The crux of the matter is whether we have or ever will have a language that is expressive enough to rigidly refer to the properties that realize $T$ without the use of rigidified descriptions. Lewis seems committed to saying “no.” The theory $T$, recall, is supposed to be the “true and complete final theory of our world.” It would thus seem that the language of $T$ is as expressive of a language as we will ever have. Thus, by Lewis’s assumption that the language of $T$ is not expressive enough to rigidly refer to the properties that realize $T$ without the use of rigidified descriptions, it seems to follow that we will never have a language that is that expressive.13

In my opinion, it is no accident that Lewis’s paper is titled “Ramseyan Humility.” Humility is Ramseyan because it’s due to a certain alleged fact about our language—namely, that our only terms for the fundamental properties that realize $T$ have Ramsey-style definitions. This, I believe, is really what distinguishes Humility from traditional skepticism.

5.4 Contextualism

It wouldn’t be right to leave our current subject without discussing the prospects of applying Lewis’s own response to traditional skepticism—contextualism—to Humility. I have chosen to discuss this type of response separately since I think it fails as a response to Humility for a reason independent of the one I’ve just been going on about.

When I gave Lewis’s analysis of knowledge above, I left out his famous sotto voce proviso. Here is Lewis’s analysis in its entirety:

**Lewisian Knowledge**  “$S$ knows that $P$ if and only if $S$’s evidence eliminates all possibilities in which not-$P$—Psst!—except for those possibilities that we are properly ignoring” (Lewis 1999, 425).

Lewis argues that, in certain contexts, we can truly be said to ‘know’ (as uttered in those contexts) that we have hands, because we are properly ignoring all the possibilities in which we don’t have hands that our evidence does not eliminate (e.g., the possibility of being a handless-brain-in-a-vat).

Can we make this sort of response to Humility? Again, Schaffer thinks so: “[Just] as the contextualist allows that claims to know that one has hands count as true when skeptical scenarios are not salient, so she should
allow that claims to know which properties exist count as true when quid-dity-swapping scenarios are not salient” (Schaffer 2005, 23). Rae Langton (2004) agrees: “if Lewis shows us how we can ‘properly ignore’ the skeptical possibility, perhaps the same strategy could show us how we can ‘properly ignore’, for example, the possibilities of permutation, or replacement by ‘idlers’, that the argument for Ramseyan Humility exploits. The good news for my reader, then, is that your knowledge of things in themselves may be safe—at least as far as the argument for Ramseyan Humility is concerned” (Langton 2004, 134–135).

I don’t think that a contextualist response to Humility has much plausibility. Lewis’s brand of contextualism, in particular, is committed to this being the case. Here is Lewis’s explanation of why we do not know that a given lottery ticket will lose, no matter how low the odds (a version of the so-called lottery paradox):

Suppose one possibility saliently resembles another. Then if one of them may not be properly ignored, neither may the other . . . . It is the Rule of Resemblance that explains why you do not know that you will lose the lottery, no matter what the odds are against you and no matter how sure you should therefore be that you will lose. For every ticket, there is the possibility that it will win. These possibilities are saliently similar to one another: so either every one of them may be properly ignored, or else none may. But one of them may not properly be ignored: the one that actually obtains [by Lewis’s Rule of Actuality]. (Lewis 1999, 429–430)

(Lewis also uses the Rule of Resemblance to explain why the subjects of Gettier cases do not know what they truly and justifiedly believe.)

If Lewis is right that we may not properly ignore possibilities that saliently resemble actuality, then it is difficult to see how there could be a context in which we are properly ignoring alternative realization scenarios. As before, suppose R1 is the actual realization of T, and R2 is just like R1 except that the first two members of R1 have been permuted in R2. What could more saliently resemble actuality (the possibility that R1 realizes T) than the possibility that R2 realizes T? This case seems exactly parallel to the lottery case. Suppose the winning ticket is T1 and not T2. Lewis explains that we cannot ignore the possibility that T2 is the winning ticket because it saliently resembles actuality (the possibility that T1 is the winning ticket). Since the possibilities here merely differ over which ticket realizes the winning-ticket role, and this seems to explain their salient similarity, I don’t see why we shouldn’t say the very same thing about the possibilities of R1 and R2 realizing T.

I don’t think that the argument against appealing to contextualism in the case of Humility applies only to Lewis’s particular version of
contextualism. It seems to me that any version of contextualism that gets the lottery case right will be useless against Humility. The trouble is simply that, as long as we’re talking about realizations of T, at least some realization of T other than the actual realization is going to be salient.14

This concludes my discussion of the responses to ordinary skepticism that won’t work as responses to Humility.

6 Two Dogs That Might Hunt

In the preceding sections we have seen that three of the most popular responses to ordinary skepticism are unsuccessful when it comes to Humility. I will now briefly discuss two responses that do have some initial plausibility as responses to Humility.

But first one qualifier. My intension here is not to defend these responses as successful responses to Humility. To do that, I would need to defend the principles upon which they, respectively, rest. My point here is simply to show that if one of these principles is right, then Humility is false (or, rather, we have no reason to think that it is true).

6.1 Antiquidditism

The first response to consider is that of denying the metaphysical assumption upon which the argument for Humility rests, namely, quidditism. If, contrary to quidditism, possibility is not preserved by freely permuting properties with respect to their nomological/locational roles, then Lewis’s claim that any permutation of T is a possible realization of T is false, and so Lewis’s argument for Humility is unsound. Since this way of responding to Humility has been discussed at length in other places, I won’t spill any more ink over this (difficult) issue here. However, since many take the epistemic consequences of quidditism—namely, Humility—as reason to reject quidditism, I hope to have shed some light on this issue by getting clearer on just what Humility amounts to and whether it follows from quidditism.

6.2 Direct Realism

The other response to Humility that has at least some initial plausibility is direct realism. Lewis himself considers what I believe qualifies as a direct realist response to Humility:

A friend of phenomenal qualia might speculate that all the actually instantiated fundamental properties are qualia. That would not preclude them from also occupying
physical roles—a sort of panpsychism. So, even if our true final theory is a physical theory, they might be the members of its actual realization. He might also accept the Identification Thesis: anyone acquainted with a quale knows just which property it is. Now it may seem that we can know the identities of the fundamental properties after all—we need only become acquainted with them. (This volume, 217)

Interestingly, Lewis rejects this response to Humility not because it requires that the fundamental properties be qualia, but because he rejects the Identification Thesis.

I do not want to debate the Identification Thesis here. My point is simply to note that (A) the response Lewis is considering seems to be something of a direct realist response and (B) the direct realist response is at least an appropriate response to Humility.

Why is the direct realist response an appropriate response to Humility? The direct realist will deny Lewis’s assumption that the O-language refers to the fundamental properties only by means of descriptions of their roles. Consider the fundamental property referred to by the T-term ‘t₁’. Now imagine the direct realist becoming directly acquainted with t₁ (as a quale) and, according to the Identification Thesis, thereby knowing which property she is acquainted with. Now imagine her “baptizing” this property with the O-term ‘o₁’. This is enough to block Lewis’s arguments that a permutation of this property within the actual realization of T is also a possible realization of T. Recall that that argument assumed that (1) the O-language refers to fundamental properties only by means of role/locational descriptions and (2) the O-language suffices to describe all possible observations. If the direct realist is right about this case, then it seems that ‘o₁’ does not refer to o₁ by means of a role-description. Thus, if the term ‘o₁’ is made part of the O-language, then (1) is not true. If, on the other hand, the term ‘o₁’ is not made part of the O-language (because, say, terms of direct acquaintance such as ‘o₁’ can only be part of a private language), then (2) is not true, because there is a possible observation that the O-language does not describe—namely, that of becoming directly acquainted with o₁. (As Lewis notes, this second possibility is rather dubious: there seems to be no reason why agents would have a private language that was more expressive than the language of T, which is, recall, the true and complete final theory of our world.)

7 Capitulation

The final response to Humility we should consider is capitulation. This is of course Lewis’s own response. He asks, rhetorically, “[Why] should I want
to block that argument? Why is Humility ‘ominous’? Who ever promised me that I was capable in principle of knowing everything?” (this volume, 211). If there is one thing that all parties to the dispute over Humility are agreed to, it is that Humility is probably the most benign form of ignorance one could have. Not knowing which properties realize T is nothing, to say the least, like not knowing that one has hands, or not knowing that one is talking to other people.

What has been less than clear, I think, is exactly why Humility is less ominous than ordinary skepticism. I think that things here have been confused because there are two ways in which Humility is less ominous than ordinary skepticism. The first, and fairly well-recognized way, is that Humility claims that we will never know something that we never believed we did know (consider again the deductionist response from above). The ordinary skeptic, on the other hand, claims that we do not know things that we certainly do believe we know.

However, it seems to me that there is a second, and perhaps more important way in which Humility is less ominous than ordinary skepticism. A full discussion of this reason will have to wait for another time, but for now, let me just give a taste. Ignorance can be ominous either because knowledge is good for its own sake, or because knowledge is useful. Insofar as knowledge is good for its own sake, then there’s just no getting around it: Humility implies that we will forever lack whatever intrinsic value there is to knowing what Humility says we will never know. However, it seems that Humility (unlike traditional skepticism) is in no way ominous in the other sense: the knowledge denied by Humility is utterly and completely useless knowledge.15 The reason for this is the same as the reason why Humility is immune to the standard responses to traditional skepticism: Humility postulates ignorance of propositions that can be expressed, according to Lewis, only via the use of rigidified descriptions. Given the nature of the decision problems we face—in particular, given that we never differentiate actions or differently value outcomes that can be distinguished only via rigidified descriptions—it seems that learning the proposition that Humility claims we will never know is, in principle, quite useless.

Acknowledgments

I would like to thank Jim Joyce, Allan Gibbard, Andy Egan, Ivan Mayerhofer, Jonathan Schaffer, David Braddon-Mitchell, Robert Nola, Aaron Bronfman, Boris Kment, Howard Nye, Shamik Dasgupta, Elka Shortsleeve, and all the participants in the 2007 NYU/Columbia Graduate Student
Philosophy Conference and the 2007 Southeast Graduate Philosophy Conference, for their thoughtful comments on earlier drafts of this essay.

Notes

1. In this stage, Lewis actually gives three arguments for the claim that T has multiple possible realizations. For the sake of simplicity, I present only what seems to be the least controversial argument here—the so-called permutation argument.

However, I would like to briefly point out what seems to be a mistake in another of Lewis’s argument—namely, the replacement-by-idlers argument (this volume, 212–214). An idler is a property that is instantiated in our world yet “plays no role in the workings of nature.” As such, idlers will not be mentioned by the theory T. Lewis gives an argument similar to the permutation argument, but which involves replacing members of the actual realization of T with idlers (rather than permuting them) to form a distinct possible realization of T. Lewis claims that as a modal realist, he has plenty of reason to believe in the existence of idlers, but he admits that actualists have little reason to believe in them. Nevertheless, Lewis asks

But how much reason has he [the actualist] to disbelieve in idlers? I grant that Occam’s razor justifies him in having a low degree of belief in idlers. But can he claim to know that there are no idlers? I think not. To say they don’t exist because we can have no reason to believe they do exist seems nothing better than an appeal to verificationism.

The upshot is that whatever you think about the metaphysics of modality, you should at least agree that for all you know, there may be an abundant supply of idlers. So for all you know, the replacement argument using idlers may succeed. But that is good enough. Humility is, after all, a thesis of irremediable ignorance. If we are irremediably ignorant about whether the replacement argument using idlers establishes Humility, then either the argument does establish it, and it is true; or else the argument does not establish it, but we cannot know that. In both cases alike, we are irremediably ignorant about the identities of the fundamental properties that figure in the actual realization of the true final theory (this volume, 213–214)

The move at the end seems to be a mistake: even if we cannot know that the replacement argument does not establish Humility, it does not follow that we are irremediably ignorant about the identities of the fundamental property that figure in the actual realization of the true final theory. If there are no idlers, then, setting aside the other arguments for the moment, there is only one possible realization of T—regardless of whether we know that there are no idlers and so regardless of whether we know that there is only one possible realization of T. But if there is only one possible realization of T, then any world where T is realized is a world where the actual realization of T realizes T. Thus, since Lewis allows that we may come to know that T is realized, it follows that we may come to know that the actual realization of T is the realization of T.

The trouble with Lewis’s argument is that he seems to be assuming some sort of “KK principle”—i.e., that knowing that P implies knowing that one knows that P. This principle holds on some accounts of knowledge, but not on others. Interestingly enough, this principle does not hold on Lewis’s own account of knowledge
(see below), since on that account what one knows depends in part upon what possibilities there are (e.g., whether there is possible world where T is realized by some property that is actually an idler) regardless of whether one knows that these are the possibilities.

2. My definition of ‘quidditism’ diverges from that of Robert Black (2000), which is the definition cited by Lewis in “Ramseyan Humility.” Black defines ‘quidditism’ as the thesis that “there is primitive identity between fundamental properties across possible worlds” (Black 2000, 92). However, as Lewis says, quidditism is to be to properties as haecceitism is to individuals. Accordingly, I have constructed my definition of ‘quidditism’ to be analogous to Lewis’ definition of ‘haecceitism’ in (1986a), where he was at pains to not identify haecceitism with the thesis that “there is primitive identity between individuals across possible worlds”. I believe my definition also squares better with the remainder of Lewis’s discussion of quidditism in “Ramseyan Humility” (see in particular this volume, 209–212).

3. I here allows Lewis’s contextualist sotto voce proviso (“Psst!—except for the possibilities that we are properly ignoring”) to remain sotto voce. As I argue below, the appeal to contextualism won’t help in blocking the argument for Humility.

4. Here and henceforth I follow Lewis in assuming that propositions are sets of possible worlds (or, more generally, sets of possibilia). No doubt much of what I say in this paper would need to be revised to take on the assumption that propositions are more finely individuated.

5. The ‘nonambitious’ qualifier on two-dimensional semantics comes from Soames 2005. The two-dimensionalism appealed to here is nonambitious in that it only claims that some statements (in particular, those containing the term ‘actually’) have two intensions (please see the next note).

6. The labels ‘C-intension’ and ‘A-intension’ are the ones used by Frank Jackson in his particular version of two-dimensional semantics (see, e.g., Jackson 1998a). In other versions, different labels have been used. My choice to use Jackson’s labels should not be taken as an endorsement of his particular view of two-dimensional semantics. Indeed, nothing more than the relatively conservative two-dimensionalism of Davies and Humberstone (1982) is required for what I say here.


8. I should note that Whittle does believe that some epistemic conclusion follows from Lewis’s argument—namely, that we do not know the “intrinsic natures” of the properties that realize T. I object to this reading of Lewis’s argument on the grounds that fundamental properties do not have intrinsic natures, and thus there simply isn’t anything there to know. But of course this subject surely needs more treatment than I can give it here.
9. However, I do believe that Schaffer is right in that one of the responses to external-world skepticism (direct realism) does stand a chance at refuting Humility. I will discuss this potentially successful response in section 6.

10. When I speak of ‘the mass role’, I mean the role of being the $i$th member of the realization of $T$, where ‘mass’ is the $i$th $T$-term in the postulate of $T$. If ‘mass’ is not a $T$-term of $T$, then Schaffer’s example will need to be adjusted accordingly.

11. This isn’t circular, just shorthand. Please see the previous note.

12. We can of course distinguish the hypotheses in trivial ways such as the proposition expressed by such and such sentence vs the proposition expressed by such and such other sentence. But I take it that those distinctions are irrelevant to which is the better explanation.

13. For my part, I don’t yet wish to take a take a position on this assumption. My point here is merely that (most of) the traditional responses to skepticism will not work as responses to Humility unless they are coupled with responses to this semantic thesis.

14. What if we’re not talking about realizations of $T$? I admit that there might be some contexts in which we are properly ignoring possibilities in which $T$ has some alternative realization. However, as soon as we even mention Lewis’s thesis, we are no longer ignoring those possibilities and so it is true. It might be responded that the contextualist response to traditional skepticism works in the same way: as soon as we mention the traditional skeptic’s thesis that we do not know that we are not brains in vats, the possibility that we are brains in vats has become salient and so the skeptic’s thesis is true. However, the traditional skeptic also claims, for example, that we do not know that we have hands. But just mentioning that thesis does not put us in a context in which the possibility that we are brains in vats has become salient. Thus, the traditional skeptic sometimes makes claims that are refuted by contextualism when he is making them. Not so in the case of Humility.

15. Again, this isn’t to say that it isn’t intrinsically valuable to have such knowledge.
Imagine a very fine grid or graph on which dots are placed at various coordinates so that, as a consequence, this or that shape materializes there. Depending on the coordinates of the dots, different shapes will appear, and for every shape there will be a pattern in the coordinates that guarantees its appearance. Take, for example, the diagonal line that slopes rightward and upward at an angle of 45 degrees from the origin. This line is bound to make an appearance so long as the coordinates satisfy the condition or pattern that as they move away from the origin, \((0,0)\), the coordinates are progressively larger pairs of equal numbers: \((1,1)\), \((3,3)\), and so on.

In the world of such dots and shapes, it is going to be in principle possible, for any array of dots that realizes a relevant shape, to derive the presence of the shape from the numerical coordinates of the dots. More particularly, it is going to be possible to derive that shape without reliance on anything other than, first, the empirical fact that the given array of coordinates instantiates this or that pattern; and second, the a priori knowable fact that the pattern guarantees the presence of the shape in question. The nature of the shapes on any grid—if indeed there are any relevant shapes present—is going to be a priori derivable from the positions of the dots; it is going to be possible in principle to derive the one from the other.

The simplest and most appealing version of physicalism parallels this sort of doctrine about dots and shapes (Pettit 1994, 1995). It holds that just as the positions of the dots determine the nature of the shapes a priori, so the way the natural world is physically organized a priori determines the way it presents itself in psychological and other terms. The way things are physically configured entails the presence of psychological and other realities, and it does this without reliance on anything other than what a
priori analysis can in principle reveal. There is an a priori entailment from the way things are physically—however ‘physical’ is understood (Pettit 1993b, 1994, 1995)—to the ways they are in other respects (Chalmers 1996; Jackson 1998a; Chalmers and Jackson 2001).

I am not going to defend my understanding of physicalism in this essay, nor my commitment to the truth of physicalism, so understood (for more see Pettit 2003, 2004). Rather I want to focus on a general problem it must confront. This is that even if such a physicalism is true, it is not going to be very satisfying in a range of important cases. I have discussed this problem elsewhere for the case of consciousness in particular (Pettit forthcoming); here I look at it in a more general and systematic way.

The problem that I raise for the physicalist derivation of psychology is that although we may be in a position to believe that the psychological is a priori derivable from the physical, so that the presence of this or that psychological phenomenon is a priori derivable from how things physically are, we are very unlikely to be in a position actually to conduct a derivation or even to get a sense of how it would go. We suffer from a derivational deficiency that takes away from the satisfaction that derivations generally give us.¹

Conducting a derivation means surveying and endorsing the premises, developing an insight into why they necessitate the conclusion, and being inclined on that basis—being thereby moved or pushed or forced—to assent to the conclusion. But it turns out that you may have excellent reason for believing that a conclusion can be derived from certain premises, and you may be able to survey and understand those premises, without the belief in the premises providing any insight-based inclination to believe the conclusion. You will believe the conclusion if you are rational, but the belief will not be the spontaneous product of understanding and embracing the premises. It will materialize without a motivating insight into why the premises make the embrace of the conclusion unavoidable.

I think that as things currently stand, we are in this position with regard to the derivation of many aspects of the psychological from the physical. Ronald Knox claimed to want an argument for God’s existence that would bring him to his knees at the conclusion. Most physicalists of my ilk would like something of that kind with mental phenomena: a derivation from physical premises that would make salient and inescapable just how such phenomena can materialize in arrangements of purely physical stuff. The thesis of this essay, however, is that at least in many cases they are unlikely to feel anything like this level of satisfaction in the best accounts—the best, by my lights—that are currently available.
The problem, as the title of the essay signals, is that we lack an experience of inferential pop-out. Most of us are familiar with the pop-out displays that look at first to be just a jumble of colors and shapes but that can come—perhaps immediately, perhaps only after attentional effort—to present a prominent gestalt to the eye. Some derivations deliver an inferential analogue of this perceptual pop-out, others fail to do so. The physicalist derivation of psychology is one of those derivations that fail, for reasons we shall be examining, to provide this experience of pop-out. And it is this failure that makes physicalism a somewhat less than fully satisfying doctrine.

The essay is in three sections. In the first, I go back to the analogue of the grid with the dots and shapes and identify two cases where we might have good reason to believe in a priori derivability without actually having the corresponding derivational ability. And then in the following sections I look in the light of those two models at the physicalistic derivability of two sorts of psychological state. In the second section I consider psychological states that are representational but not recursively representational, as I shall put it. In the third I investigate the rather more complex and troublesome case of states that are recursively representational. These are states such that not only do the subjects of the states represent things to be thus and so; they are also in a position to represent things as being so represented by them.

1 Two Models of Derivational Deficiency

1.1 The Diagonal Case

Returning to the grid envisaged earlier, it is relatively easy to show how a derivation might be provided for the presence on the grid of an upward, rightward line that starts at the origin and slopes at an angle of 45 degrees. Call this line the diagonal slope and call the pattern that ensures the presence of such a line, pattern D; this is the numerical pattern under which the first coordinate is (0,0) and others are progressively larger pairs of equal numbers: they take the form (1,1), (2,2), . . . , (m,m). . . . The derivation will go like this:

(1) A priori: any coordinates in pattern D generate the diagonal slope.
(2) Empirically known: the coordinates plotted on the grid instantiate pattern D.
(3) Conclusion: the coordinates plotted generate the diagonal slope.

The first premise in this argument gives us the design specifications for the diagonal slope, as we might put it, and the second premise asserts that
they are satisfied. You can easily tell the design specifications—the recipe for constructing a diagonal—from the definition of the diagonal slope, recognizing that only dots with coordinates in pattern D will lie on that slope. And you can tell that the coordinates conform to that pattern by ready inspection. Not only will you be able to see, then, that the conclusion is derivable from the empirical information, without reliance on anything other than a priori presupposition. You will be able to conduct that derivation with little or no trouble from within your own, internal perspective. As soon as you survey, understand, and endorse the premises, you will be moved by the insight they provide to endorse the conclusion. There will be nothing to give you pause.

The inference may come to you so naturally that you can be said to see the diagonal slope in the sequence of number-pairs; you may even see it there as surely and determinately as in the visual appearance of the graph itself. The number-pairs will contain the information that the dots are on that slope, since the connection with the slope is a priori. And, even more strikingly, they will impart or convey that information in a way that will move you spontaneously to admit the conclusion.

Your position in this case may be so good, indeed, that you barely notice the numbers when you register the coordinates; you just register straight-away the diagonal slope that the numbers entail. Think by analogy of how the accomplished performer can look at the score of a piece of music and hear the profile that that score determines. Or about the doctor who can look at the shadows on an X-ray and see the ulceration or the cyst that completely escapes the unpracticed eye. Or about the sailor who can look at the water in the harbor and see where the various currents and rips are running.

I have been concentrating on the ease with which you can conduct the derivation of the diagonal slope in order to emphasize the element of skill—inferential habituation—that is needed for this exercise. But even if the ease of inference is not so great, we can still think of you as conducting the derivation and as seeing the diagonal slope in the coordinate pairs. It may take you some time to work out that a given array of coordinates satisfies a certain pattern and it may come to you only slowly, say as a result of going through various tests, that this is the pattern for the diagonal slope. But it still remains that at the end of the exercise you will be inclined on the basis of insight to assent to the presence of the slope; you will find yourself compelled to countenance the slope. The inference from the array of coordinates to the nature of the slope will come upon you with all the irresistibility of an experience of pop-out.
1.2 Shallow Derivational Deficiency

Let us move now from the simple case of the diagonal slope to a figure that is not so easily definable and not so familiar: that of an S shape. Just as there will be design specifications for a diagonal slope that can be worked out a priori, so the same will be true with this shape. Starting from the bottom, an S consists in a line that, within certain degrees of tolerance, moves down slightly and slowly to the right for a certain distance; then turns upward and leftward sharply for about twice that distance; and so on. Knowing this—knowing the recipe for constructing an S shape—you will be able to work out by analysis that only dots with coordinates in a certain pattern—call it pattern X—will constitute an S shape. You may have to do this by looking at the different shapes within a given scale that would conform to your sense of an S shape, answering to the demands of your perceptual concept, and then working out the constraints on the coordinates of any such S-type shape. Or you may be able to do it by finding an abstract, algebraic formula for constructing an S shape and then working out the limits on coordinates that will satisfy that formula.

Given that this is the case, you will be able to endorse an argument for the a priori derivability of the presence of an S shape of the following sort.

1. A priori: any coordinates in pattern X generate an S shape.
2. Empirical: the coordinates plotted on the grid are in pattern X.
3. Conclusion: the coordinates plotted generate an S shape.

Here, as in the other case, the number-pairs presented contain the information that the dots form an S shape. But unlike the other case, they will not spontaneously impart or convey that information to you; they will not move you more or less spontaneously to admit the conclusion. The reason for this is that though the two premises are both within your epistemic reach, they are not at all obvious. It may not be obvious that any dots with coordinates in pattern X will make an S shape, however a priori that truth. And it certainly may not be obvious that the coordinates given conform to pattern X.

The reason for this lack of salience in the premises, quite simply, is that S shapes are much less commonly confronted in using geometrical displays than are diagonal slopes. No doubt, if you paid regular attention to constructing S shapes—the sort of attention you may pay to diagonal slopes—then you would develop a sense of the recipe for making an S shape, and you would get into the habit of seeing the required pattern X in sequences.
of number pairs. And in that case you might become as habituated in the inference from coordinates to shape as in the other case.

Given that you lack that sort of familiarity with constructing S shapes, however, you will not be derivationally moved in this case in the manner in which you were moved in the last. You may have reason to accept the soundness of the argument given, and thereby register that the presence of the S slope is a priori derivable from the nature of the coordinates. But you will not be in a position to conduct the derivation: you will not be spontaneously susceptible to the reason the premises provide for endorsing the conclusion.

What this case shows, then, is that there can be a gap between knowing that a conclusion is a priori entailed—this is knowledge possessed here as in the diagonal case—and having the derivational skill or know-how whereby you can be moved by understanding and endorsing the premises into drawing the conclusion. In this case you are subject to a derivational deficiency that blocks the experience of an insight-based inclination to assent to the conclusion. It blocks the possibility of inferential perception and pop-out, to return to the metaphor introduced in the previous discussion.

The deficiency illustrated is shallow rather than deep, coming from a shortfall in the level and quality of attention you pay to how to construct S shapes. Thus it is a deficiency that can be more or less easily put right. You won’t need any new sort of skill in order to develop a facility in deriving the S shape of the kind that you enjoy with the diagonal slope. You will only need to extend familiar skills into this new domain, mastering the S recipe as you mastered the recipe for making a diagonal.

There are many cases where lack of familiarity gives rise to the sort of derivational deficiency that we have been illustrating. Think of the beginner pilot who knows exactly what the instrument readings are telling her but who struggles to let them override her kinesthetic sense: who struggles, say, to conclude that she is accelerating when her kinesthetic intuition, without a view of the horizon, is that she is falling. Or think of the inveterate gambler who learns of the gambler’s fallacy—the fallacy of thinking, say, that a run of heads with a fair coin makes tails more likely on the next throw—but who finds it very hard indeed to apply this knowledge in practice. In both of these cases, as in many more, there is a theoretical form of inferential knowledge but a lack of practical inferential skill (Pettit 1998; McGeer and Pettit 2002). In both cases there is a failure of the person to perceive or to be primed by the pattern in the data. The person believes
that that pattern is there, and has excellent reason for believing it, but does not find it inferentially salient or stimulating.

The fact that you can actively derive the presence of the diagonal slope from suitable coordinates—the fact that you can see the slope pop out from the coordinates—means that the claim about the derivability of the slope is going to be intuitively or phenomenologically satisfying. By contrast, the claim about the derivability of the S shape is going to be less satisfying; it asserts a relationship that you accept but cannot grasp in quite the same way. What will be available is only the proxy satisfaction of knowing from analogous cases like that of the diagonal slope what it would be like to be able to effect the derivation. You will be able to imagine the experience of being moved by insight into a pattern in the coordinates to acknowledge the presence of an S shape. You will be like the beginner pilot, or the gambler who just cannot internalize and apply the lesson of the gambler’s fallacy.

The position you are likely to be in with the S shape has parallels in the case of visual pop-out. You may know what it is for a gestalt to pop out of a visual display, and may know that there is a gestalt of, say, a cube to be seen in a certain pointilliste display. But you may just not be able to see the cube in the display—unlike others, perhaps. You may have every reason to believe in the presence of the cube, and you may understand what this means from experience with other pop-outs, but you will still fall short of the full, phenomenological satisfaction that goes with seeing it there. Your position with the S shape will be exactly parallel. You will know what it would be like to be able to derive the presence of the S shape, as a matter of inferential habit and compulsion, but you will not be able to enjoy that inferential experience yourself. You will be rationally required to acknowledge the presence of the S shape but you will not have any sense of inferential perception.

1.3 Deep Derivational Deficiency

Sticking with the general analogue of the grid and the dots, I now introduce a novelty that is designed to illustrate the possibility of a deeper derivational deficiency. In the cases considered so far, both plotted coordinates and the shapes they make are registered visually from a close-up consideration of the grid. Thus, by going back and forth between our visual tracking of the coordinates and our visual sense of what will count for us as a given shape—this, whether or not we have a formula for the shape—we can see that there is a necessary connection between the coordinates satisfying a certain pattern and the shape being present; we can see the design
specifications for the shape—the recipe for constructing the figure—expressing these as demands on the coordinates.

Consider now a case where the coordinates determine a shape, as before, but also present a gestalt or profile, as of a three-dimensional object, when you stand back from the grid. The profile is a distally discernible property of the shape. It is something that can be seen in the shape but only when you stand at a certain distance. The man-in-the-moon is a profile in this sense. And so is the figure that emerges in certain impressionist paintings, when you stand back from them. As the coordinates on a grid may fix the presence of a regular, proximally discernible shape, so clearly they may fix the presence of such a distally identifiable profile. The same argument will apply in both the simple shape-case and the profile-in-the-shape case.

The profile introduces a possible limitation that is very unlikely to strike in the other case. In the other case, whether it be that of the diagonal line or the S shape, it is possible to go back and forth between plotted coordinates and shape, since they are observable from the same viewpoint, to see how the pattern of determination or dependency goes, and to learn the recipe for constructing the shape. You will see the pattern in salient exactitude with the diagonal line; you will see it only in broad outline—broad but scrutable outline—with the S shape. But in the profile case nothing like this may be possible, for you may not be able to go back and forth in the same way, tracking specific dependencies of profile on coordinates. Look close enough to detect the plotted coordinates and you won’t see the profile. Go far enough away to see the profile and you won’t be able to detect the coordinates plotted.

Imagine in addition that though you can change visual standpoint in relation to a grid, you can’t rely on memory or notes to track the sorts of dependencies that obtain between a profile that appears at a distance—say, the profile of the pope’s face—and the coordinates of the constituent dots. You can’t go back and forth in the way you could with the ordinary shapes, and you can’t go back and forth in memory or imagination. You can’t master the recipe for constructing the profile, then; you will lack the epistemic resources that mastery requires.

A deep derivational deficiency will arise, if despite this limitation, you have good reason to believe that the coordinates on the grid—and any of a range of coordinate-sets satisfying a similar pattern, \( Y \)—allow the a priori derivation of the profile. And of course you will have reason to believe this, given the parallel between the profile case and the case with the regular shapes. You will have reason to think that there is an argument available of the usual kind:
(1) A priori: any coordinates in pattern Y generate a papal profile.
(2) Empirical: the coordinates plotted on the grid are in pattern Y.
(3) Conclusion: the coordinates plotted generate a papal profile.

When you have access to the argument given, but only under the limitation mentioned, then you will suffer from a deeper derivational problem: than in the shape case. As with the S shape, you won’t be able to detect easily if the Y pattern is present. But where you could at least have a sense of the recipe for constructing an S shape, and so of the sort of coordinate pattern that will ensure the presence of an S, you won’t have any way in this case of achieving similar insights. You will not be able to determine the design specifications for the papal profile, expressed in terms of coordinates. There will be specific dependencies of the profile on the coordinates but these will not be salient, as in the diagonal case, nor scrutable, as in the case of the S shape; at most you will be able to venture only a broad hypothesis as to the form they take.

The effect of this deeper derivational deficiency on your sense of the a priori connection between the coordinates in pattern Y and the papal profile will be dramatic. It will mean that from your point of view there may be something quite inscrutable about the fact that the Y pattern guarantees the presence of the papal profile. You will lack the sense of understanding that would come with being able to see how in particular the profile depends on the coordinates.

This being so, we should notice that you are liable to have illusions about how the coordinates might be varied and the profile still preserved. You will have a sense of the profile—an ability to recognize and imagine it, for example—that is not tied to registering the coordinates satisfying any pattern in particular; and vice versa. And so, although you believe in an abstract manner that there is a pattern in the coordinates that guarantees the presence of the profile a priori it may seem that you can imagine the coordinates and the profile varying independently. You won’t have the concrete, working sense of the dependency of profile on coordinates that would banish such illusions of imaginability and conceivability.

The shallow derivational deficiency will block you, as we saw, from being able to experience anything like inferential pop-out, though you will at least have a sense of what such pop-out would be like. You will be in the position of someone who is familiar with visual pop-out and knows there is a figure of a certain kind present in a certain visual display, but who just cannot get to see that figure. The deep derivational deficiency we have been discussing will make for worse trouble and more dissatisfaction. It
will mean that not only do you not experience inferential pop-out, you will not have a sense even of what such pop-out would be like.

With our two versions of derivational deficiency distinguished, we have two models of how one may be in a position to believe in a priori derivability without having the ability actually to conduct a derivation. It is time now to apply those models to cases where physicalism claims to support the a priori derivability of psychological states. I shall argue that shallow derivational deficiency is the only problem in the first case but that the deep derivational deficiency also affects the second.

2 Nonrecursively Representational States

The information about the coordinates of the dots in our examples is parallel to the neuronal—better perhaps the neuro-environmental—information we might have about a psychological subject. If the sort of physicalism I accept is true, then that neuronal information will contain information about the psychological states of the subject, just as the information about the coordinates of the dots contains information about the shapes constituted by the dots. As there is an issue then about inferential insight in the diagrammatic case, so a similar question arises here. Can we expect as physicalists to be able to achieve—or to be able in principle to achieve—the sort of perception or insight that can prove elusive even in the diagrammatic case? Can we expect to be able to attain inferential pop-out?

I will discuss this question with regard to nonrecursively representational states in this section and then recursively representational states in the next. I will argue that our position with states of the first kind is like our position with S shapes—not ideal, though not too bad—but that our position with states of the second kind is worse again; it is like the position with the profiles coded for, given we are denied access to how coordinates can vary and create varying profile effects.

The distinction between recursively and nonrecursively representational states is relatively straightforward. Assume that representation can be naturalistically analyzed. Putting aside issues to do with explicit and implicit representation, and local or distributed realization, I propose that for current purposes we think of a representational state on the simple model that requires fulfillment of two conditions. First, that there is a generally robust connection—one that obtains in favorable conditions, however they are cast—between the form the state takes and the way the environment is configured; and, second, that the form taken by the state tends to lead the subject to behave in a manner that is intuitively appropriate for
such an environment. However it is analyzed—whether in terms of these simple requirements or in some richer fashion—there is a more or less inescapable distinction between nonrecursively and recursively representational states.

Nonrecursively representational states will involve the representation of the environment without any representation of it as being so represented. For the subject whose only representational states are nonrecursively representational, there will be no distinction between the environment as it is in itself, then, and the environment as it is for that creature—as it is, according to the creature’s representations. The subject will be primed to respond in different ways to varying scenarios in the world, assuming something on the lines of the simple account of representation sketched above. But such variations in its environment will produce those different responses without the fact that they produce them being in any way represented by the creature. The varying scenarios will be represented in the subject, as we might say, but they will not be represented for the subject (Cummins 1983). They will be represented as the ways things really are, priming the subject to adjust appropriately; this must be the case with any effective representation. But they will not be represented as the ways things are represented by the subject as being, nor a fortiori will they be represented in any further, recursive manner: they will not be represented as the ways things are represented as represented as being, and so on.

Recursively representational states, by contrast, will be states such that the subject not only sees the environment as he or she represents it to be, being disposed to adjust and act appropriately; the subject will be in a position to see the environment as being represented in that manner. The environment will be represented as real in such states, assuming they are effective representations, that is, assuming their connection with adjustment and action is not suspended, as in the case of doubt. But it will be available to be represented as a represented and possibly not real environment. And indeed the recursive availability of such representation may be open ended, with the subject having the capacity to form representations, at progressively higher levels, of how things are according to lower-level representations. The further representation will become available in each case, we may assume, at the point where it is called for by some task on hand: say, by the task involved in checking the accuracy of a representation. It will be available on a need-to-know basis.

I shall assume that there is no particular difficulty for those in any camp about acknowledging the difference between recursively and
nonrecursively representational states; more on this in the next section. It is close to the distinction drawn by Ned Block, at least in his later formulations of the idea, between representational states to which the subject has access, being able to make judgments about the way things are represented to be in those states, and representational states to which the subject has no access (Block 2002). Although there is certainly controversy as to whether consciousness involves more than access of this kind, no one appears to doubt that such access is possible with some states and not possible with others.

For an example of nonrecursively representational states, we might think of degrees of probability, and perhaps preference, as they are depicted in standard decision theory. Decision theory requires of those states only that they mutate in certain ways under various forms of evidential input and that they cohere with, and perhaps explain, the choices that the subject makes. The subject may assign a high degree of probability to a certain scenario without being aware of that scenario as something it represents as obtaining or as probably obtaining. It may assign a high degree of preference to a prospect without being aware of that prospect as something that it represents as attractive. The subject may be just a well-engineered artifact that does not itself have any beliefs about how the world is according to its beliefs, as distinct from how the world is, period; it may lack the very concept of representation.

Let us assume then that there are psychological states of a representational but nonrecursively representational kind. And let us suppose that the way things are neuronally organized in a subject instantiates representational states of this kind. If the simple account of representation works, the way things are neuronally organized will dispose the subject to adjust to circumstances—to respond to evidence and to initiate action—in an appropriate manner; and if that account fails, then something different or extra will be required.

Might we be able to see the subject’s neuronal configuration and, recognizing the pattern required for the representational states in question, just see in it the presence of those states? Seeing the subject’s neuronal configuration will mean knowing how it is neuronally constructed and how it is integrated with environmental inputs, described in neuronally relevant terms. The question then is whether we could ever be able to move smoothly from a state of registering such a neuronal or neuro-environmental pattern to a state of registering the subject’s representational profile: registering its system of representational states.

In order to achieve this insight we will have to be able to have a sense of the design specifications on nonrecursively representational states and
we will have to be able to identify relevant neuronal configurations as satisfiers of those specifications.

Under the simple account of representation sketched—or indeed under many variations on that account—it will be relatively easy to grasp the design specifications. A representation will be present so far as there is a state that covaries in form with the putatively represented situation and that disposes the agent to act as is intuitively appropriate—given the agent’s goals and collateral representations—to that situation. Thus the design specifications—the recipe for constructing a simple representational state—will impose limits on the causal sensitivity of the agent’s states to the environment and on the causal productivity of those states in generating behavior.

But though we should have no problem in getting a grasp on the design specifications for a nonrecursively representational state, we may well have a problem in identifying one or another neuronal array as a satisfier of those specifications: that is, in recognizing the presence of the pattern that can guarantee the presence of the state. We are certainly likely to have this difficulty with an entity as complex as the human being, or perhaps any natural organism. But it may well be possible to overcome this problem in the parallel case of the simple piece of artificial intelligence, robotic or otherwise. We can imagine learning the input and output susceptibilities present in the electronic or mechanical makeup of a simple robot and then, as we take those dispositions into account, recognizing them as guarantors of the presence of representational states of this or that kind. We can imagine being attuned to how the physical structure of the artifice more or less manifestly implements the representational attitudes we ascribe; such attunement would certainly be accessible if we knew how to make such a robot. We would see the attitudes present in that structure as surely as the mechanic sees the different working parts of the car engine in the complexity beneath the hood.

Looking at the coordinates for a diagonal slope on a grid, we may enjoy a sort of inferential perception or pop-out. We may be able in the same way to have inferential perception of the presence of certain nonrecursively representational states in the case of the simple artifact, or at least we may be able to get close to that sort of experience. But it is unlikely that any one of us can achieve this insight with the complex, organic subject. In this case we will be under the same sort of limitation that besets most of us with the perception of the S shape in sequences of coordinates. We will suffer from the shallow form of derivational deficiency.

Assuming that some psychological states are nonrecursively representational in the sense explained, we will be able in principle to access the following sort of inferential knowledge in this more complex case.
(1) A priori: any neuronal states in pattern Z realize such and such representational states.
(2) Empirical: the neuronal states in the brain of this agent conform to pattern Z.
(3) Conclusion: the neuronal states of the agent realize representational states of that kind.

The inferential knowledge that this argument gives us in the case of the complex subject—the knowledge of an a priori entailment that it yields—will not provide anything like inferential perception. It will not give us the ability to see the presence of the representations in the neuronal states.

This will be a source of dissatisfaction for any defender of my preferred variety of physicalism. But it need not be very frustrating. For just as there need be no deep puzzle involved in recognizing that a certain sequence of coordinates realizes an S shape—not at least for those of us who are capable of seeing simpler figures in such sequences—so there need be no particular puzzle associated with recognizing that a certain profile of neuronal states realizes a corresponding representational profile. In particular, there will be no puzzle involved for those of us who have an inferential sense of how a simpler electronic or mechanical profile can realize a simpler representational profile, knowing the sort of recipe we would follow in constructing it. Surveying such a neuronal configuration might give us no perception or insight into the representational profile that it implements but we can at least see what would be involved in achieving that sort of perception; we can draw on the parallel with the simpler case to give us a sense of this more complex counterpart.

What we can achieve in the case of relating nonrecursively representational states to neuronal realizers is a sort of reduction that is familiar in philosophy. Take the reductive thesis in social ontology—an individualist as distinct from physicalist thesis—according to which the social entities and processes that exist in any social domain can be derived in principle from the dispositions and relationships of individuals (Pettit 1993a). With social entities and processes of any complexity—say, with banks and money and market exchange—there is going to be no possibility of achieving anything more than we can achieve in the derivation of S shapes. But that need not make for a serious problem, for there are lots of toy examples in this area where we can exercise that sort of derivational skill: we can see the presence of certain simple social realities in suitable specifications of individual attitudes and interactions; for example, we can see the presence of a convention in the behavior of people who hold certain attitudes
toward the coordination of their activities (Lewis 1969). The existence of those examples makes accessible the claim maintained in more complex cases, as the example of deriving the presence of a diagonal slope makes accessible the claim maintained about the derivability of the S shape.

Daniel Dennett (1979) speaks in the sort of psychological case we have been discussing of a difference between the physical stance in which we survey the electronic or neuronal construction of a subject and the intentional stance in which its representational or intentional profiles become salient. Does the situation we have been describing merit this talk of a difference of stance? It does, so far as information is presented on the one side in neuronal or electronic terms and on the other in terms of what is (nonrecursively) represented. This scenario corresponds rather nicely, as it happens, to our model case where numerical information about the dots contrasts with figural information about the shapes.

But though there is a reason for talking here of a difference of stance, in another respect such talk is somewhat exaggerated. The reason is that numerical and figural information in the one case and neuronal and (non-recursively) representational information in the other are varieties of information that can be available simultaneously. There is no special difference of perspective or standpoint required for moving between the two. And so there is no difficulty about grasping the sort of recipe whereby, operating on neuronal materials, I might construct a certain representational pattern. Just as I am, I can absorb either the numerical or the figural information about a geometrical diagram. Just as I am, I can absorb either the neuronal or the representational information about a psychological subject. This is the point, so it turns out, where recursively representational states are importantly different.

3 Recursively Representational States

Recursively representational states are psychological states that enable the subject, not just to represent the environment after a certain fashion, but to represent it as an environment that is, precisely, represented. Recursively representational subjects may not be able to achieve recursion with every representational state they instantiate; some states, as it is said, may be subpersonal and unavailable to recursive representation. People will count as recursively representational subjects, on my usage, so far as they can achieve recursion with any representational states, however restricted the range of recursively representable states.
There ought to be nothing particularly controversial about positing the existence of recursively representational states and subjects. The recursion may obtain just in virtue of the subjects becoming able to form beliefs, not only about how the world is, but about how the world-as-represented is, or, alternatively, about how the world appears or seems to be. And there is scarcely any denying that subjects like us do form such recursive representations. The recursion need not involve any experience of the initial representation; the recursive representation may be purely a matter of belief, not of perception or sensation (see Carruthers 2000). And the belief involved may bear entirely on how the world is according to the original representation—how the world seems; it need not be an introspective or reflective belief, except implicitly, about how things are within me. We may describe the recursive representation as a metarepresentation or a higher-order representation, but it is important to be clear that what it primarily represents is the world-according-to-the-relevant-representation—if you like, the content of that representation—and not the state of representation itself.

The importance of representational recursion from our point of view is that whereas simple representational subjects will be “lost in” the world, as that world is represented by its states, recursively representational subjects will be able in principle to make a distinction between how the world is and how it is represented as being. They will do this just so far as they identify how the world is represented as being—form beliefs as to how it seems to be—and come to believe that it is not actually that way: the represented-way the world is, so they can believe, is not the real-way it is.

There is now a well-established habit of glossing the notion of an experiential or phenomenally conscious state in terms of there being something it is like to be in that state (Nagel 1986). Recursively representational states, for all I argue here, need not be experiential states in this sense. There may be a close connection between being phenomenally conscious and being recursively representational (Pettit forthcoming). But nothing in the argument defended here depends on positing that connection, or on interpreting it in any particular way.

The most obvious candidates for recursively representational states are experiential perceptions, in particular those involving perceptual illusions that can survive conflicting beliefs. Consider the Müller-Lyer illusion in which two lines of equal length look different because one line has an arrowhead at each end, the other a reverse arrowhead. Subjects who have the idea of representation can recognize that despite appearances the one line is not longer than the other, and have beliefs about how the lines are
represented for them—how they appear, what they are like—as distinct form how they are in themselves. There will be something very stable—something irritatingly stable—that the lines are like for the subject capable of recursive representation; there will be nothing that the lines are like, not at least in the same sense, for a creature without that capacity.

But recursively representational states include many representational states that are nonperceptual or at least not so purely perceptual. Consider the way things present themselves when I see them as requiring this or that response: this is the way to go on, I think, in using a certain word; this is the way to behave, I conclude, in determining my overall duty. Or consider the way the world is depicted by me when I see certain options as alternatives between which I can choose, and choose freely: these are directions in which I can move from the status quo. Or consider the manner in which I view someone who has apparently done me harm when I feel resentment and depict her as a responsible agent, fully deserving of blame. In all of these configurations of attitude—and clearly they are at the center of much of human life—the world assumes a certain form, according to my representations, and I am aware of this as a represented form that they have. I see the world in terms of certain patterns—those that go with concepts of obligation and freedom and responsibility—and I am aware of the world as represented, whether or not mistakenly, in that way.

It is probably clear where this line of thought is leading. The contrast between nonrecursively and recursively representational states introduces a difference like that between the model involving the shapes and the model involving the three-dimensional gestalts or profiles. Plotted coordinates and shapes are simultaneously available, being each available from the same visual standpoint, and so it is possible for someone who can recognize a given shape, having the concept of that shape, to work out the coordinates that will ensure its realization; it is possible for someone to get a sense of the recipe for constructing the shape. Plotted coordinates and profiles are not simultaneously available, requiring different viewpoints, and under the model we described there is no possibility of moving between them in the same way. There is no possibility of grasping the recipe for constructing such a gestalt or profile.

Routinely representational states and neuronal configurations are simultaneously available in the manner of coordinates and shapes, at least under our simple account of representation or any of a number of variations on that account. But recursively representational states and neuronal configurations are not available in the same way and, as with coordinates and
profiles, it is not possible in the ordinary run of things to move back and forth between the two.

The analogy between the profiles case and the case of recursively representational states becomes apparent once we ask after how we might identify the design specifications on a recursively representational state: the recipe for constructing such a state out of neuronal materials. It will not be enough for gaining a sense of this recipe to recognize the causal role that neurons will have to play, as this was enough with ordinary representational states. And this makes for a crucial difference between regular and recursive representation.

The recursively representational state, qua representation of the represented-way-things-are, will have to covary with variations in the represented-way-things-are, and will have to channel behavior that is appropriate in light of the subject’s desires: say, the behavior of interrogating the represented-way-things-are in order to answer this or that query. But we will not be able to register variations in the represented-way-things-are, nor identify suitable behavior in relation to the represented-way-things-are, without actually instantiating the recursive representation. With the simple representational state we merely had to observe the environment in order to plot neuronal covariations in response to variations in the represented content and we merely had to observe the neuronally prompted behavior of the subject on the environment in order to chart suitable behavior. With the recursively representational state nothing like this is possible.

It will be necessary in this case to gain access to the represented-way-things-are that becomes available in recursive representation and to correlate it with neuronal changes. But the only way of gaining this access will be by undergoing the recursive representation in question. And so the only way of achieving a sense of the design specifications for recursively representational states—the recipe for constructing such states out of neuronal materials—will be by mapping variations in neuronal configuration, as they are registered from a third person viewpoint, with variations in the represented-ways-things-are, as they are available in first-person experience.

The shift of viewpoint required here is directly analogous to the shift of viewpoint required in order to try to determine the design specifications for a profile as distinct from a simple shape. In the profile case, we have to move from the proximal position where the coordinates are visible, but not the profile, to the distal position where the profile is visible, but not the coordinates. In the case of the recursive representation, we have to
move from the third-person standpoint where the neurons are capable of being registered, but not the represented-way-things-are, to the first-person standpoint where the represented-way-things-are can be determined, but not the way the neurons are configured. As there was a deep derivational deficiency present in the profiles model, so there will be a deep derivational deficiency present in this case.

In the profiles model, it would be possible for us to gain a sense of the design specifications for the papal profile if we were able to go back and forth between the plotted coordinates that determine the profile and the profile itself. What holds here in parallel is that a deep derivational deficiency will be avoidable so far as we can go back and forth between neuronal observation and first-person experience, coming to learn, however roughly, of the way in which neuronal variation makes for variation in our own recursive representation.

But there’s the rub. The deep derivational deficiency that threatens in this case looks even more inescapable than the corresponding deficiency in the profiles case. For there is no easy analogue to the recipe for repair that we offered in that case. You can overcome the difficulty with profiles just by being given access to the connections between variations in the plotted coordinates you see and variations in the profiles you see. But there is nothing to give you access to the connections between neuronal variations and variations in how the world is recursively represented.

Or at least there is nothing to give you this access at present. It is entirely possible that future technology will enable us to gain a sense of how recursive representation depends on neuronal variation, letting us see patterns that our neuronal configurations must satisfy if they are to sustain this or that form of recursive representation. But as things currently stand, this is something that is effectively denied to us. We may have very good reason to believe, as physicalists believe, in the existence of patterns that provide an a priori demonstrable guarantee—a guarantee demonstrable in principle—of the presence of certain recursively representational states. We may even have well-argued hypotheses as to the general form that those patterns are likely to take. But we may still suffer from a deep derivational deficiency.

The effect of the deficiency, as in the profiles case, may be to deprive us of a sense of what is imaginable and conceivable that answers to our beliefs about what is a priori necessary. The sense of the conceivable and the inconceivable—the intuition into what is possible and impossible—will be responsive, plausibly, to our sense of specific dependencies. And that sense of dependencies will be missing in this case, as with the profiles. Thus,
regardless of our beliefs as to what is a priori guaranteed by certain neuronal configurations, we may think we can imagine things running counter to those beliefs. We may find ourselves with an intuition that a given configuration might be present without the corresponding represented-way-things-are, and the other way around.

Take a recursively representational state such as seeing red, or following a rule, or identifying a behavioral option, or endorsing a feeling of resentment toward someone. It may be possible for us to try to work out the sorts of functional and physical conditions—ultimately, the neuronal patterns—that would suffice for such recursive representations. But even as we try to assure ourselves that such-and-such conditions would be bound to make things present themselves the way that they intuitively present themselves in these experiences, we have to recognize that all we can do is argue and assert. We will lack the spontaneous intuitions to back up these hypotheses.

Thus, we may analyze a conscious experience such as that of seeing red in terms of information-processing (Pettit 2003), but, notwithstanding the case for that hypothesis, there remains the intuition that the physically determined processing might remain fixed while the color appearance changes or fades. Again, we may analyze the experience of rule-following in terms of how things are bound to seem from the perspective of an agent with certain extrapolative and regulative dispositions (Pettit 2002), but there still remains the intuition that these dispositions could be in place without the subject entering normative space and having a sense of the right and the wrong, the appropriate and the inappropriate. We may analyze the notion of responsibility in terms of the capacity we attribute to those with whom we think it is worth conversing, a capacity that we take them to possess even when they fail to exercise it (Pettit and Smith 1996; Pettit 2001a,b). But there still remains the intuition that the capacity we attribute requires nothing more than complexity—in particular, a complex susceptibility to conversational influence—and that it could well be in place without the freedom of the will that responsibility strictly presupposes.

Were we able to have a concrete sense of dependencies in these cases, rather than abstract hypotheses as to the form they take, then we would be able to see the sort of recipe whereby a recursive representation might be constructed out of neuronal materials. And so we would be able to experience something like inferential pop-out. Or at least we would be able to have a sense of what such pop-out would be like. We would be able to simulate the fulfillment of the antecedent conditions given in the analyses,
and find ourselves driven to acknowledge that under those conditions, the represented-ways-things-are would be bound to fall in line. But in all these cases effective simulation will be denied us, lacking as we do the opportunity to track the specific dependencies of recursive representation on neuronal pattern.

Hume is well known, at least under some interpretations, for arguing that the experience of having one sort of event more or less invariably precede another leads us, mistakenly, to posit a necessary connection—by his account, a causal linkage—between the first and the second types of occurrence. The derivational or projective fallacy that he imputes to us is the inverse of the derivational or projective deficiency alleged here. In his case, there is no necessary connection between the distinct events envisaged, and the problem is that there is an experience of being led by the occurrence of events in the first category to the expectation of the occurrence of events in the second. In our case, there is a necessary connection between physical and psychological phenomena—an a priori entailment—and the problem is that that there is no experience of being led by variations in the first domain to the expectation of variations in the second. It is this that puts the experience of pop-out beyond our reach.

Hume’s problem is that our psychological habits induce us to posit among things connections that do not exist there, whereas my problem is that there are connections among things that our psychological habits do not induce us to posit. He rails against the siren songs of nature and custom; I lament the failure of nature and custom to sing. There is an allegation on both sides that mind and world are misaligned at the level of epistemic impulse, but where he indicts epistemic impulse as a source of error—it indicates the presence of a nonexistent relationship—I indict it as a source of ignorance: it fails to signal the presence of a relationship that does by our account obtain.

One final, consoling thought. The derivational deficiency that afflicts us according to the story told here may eventually prove to be remediable. Take the physicalistic claim that the way colors recursively present themselves is determined by how subjects are primed to register and process color-related information—information to do with illuminance, contrast, constancy, and the like (Pettit 2003)—using it as a basis for accommodating to their environment. At least one experiment shows—and, in the relevant way, shows the subjects of the experiment—that when the informational process is disturbed, then color-perception is disturbed too, and that when the informational process is restored so as to ensure smooth functioning, then color-perception is restored at the same time. This
amounts to showing an effect in how the world is presented visually to a subject—an effect in how colors are recursively represented—that is consequent on physical changes in his or her makeup.

The changes induced in the subjects of this experiment were brought about externally. The researcher had them wear glasses in which the lenses were colored in different ways—red at the top and green at the bottom, or blue on the right edges and yellow on the left (Kohler 1961, 1964). The idea was to see whether the way colors presented themselves would change—whether the distortions would disappear—as the brain adjusted and leached them out: that is, as it extracted information from the environment, consistent with the background, perhaps hardwired assumption that things do not change color when one moves one's head. The finding, exactly as a physicalist would predict, was that as soon as the visual system got back to processing color information consistently with that assumption, and in a manner suited to smooth color-based discrimination and behavior, the way things looked for the subject did indeed change; it returned to normal (Pettit 2003).

This adjustment took a long time to materialize—a matter of several weeks—and involved a good deal of inconvenience. But imagine that one could induce the changes in minutes or even hours and see for oneself how the normal color appearance goes with normal adjustment in the registering and processing of related information. In that case one would surely be within reach of pop-out. One would be able to bring oneself to the point of vindicating the physicalist analysis of color appearance in terms of an ability on the part of the subject of the appearance to resolve and make use of certain color-related information. And one would be able to get a vivid inferential sense of how, in the presence of the appropriate, physically based ability, it is impossible that the subject not enjoy color appearances, and not enjoy precisely these or those appearances in particular. I see no reason, then, to despair. The frustrations of physicalism may not be with us forever.

4 Conclusion

Physicalists of my preferred stripe have to think that there is no room for conceiving of a gap—not at least under full information and understanding—between things being physically thus and so and this or that psychological state materializing. They have to think that once the physical is suitably fixed, then as an a priori discernible matter the psychological will be fixed as well. But that commitment is hard to sustain because of a variety
of intuitions to the effect that certain recursively representational phenomena can vary independently of the physical.

I think that there are good arguments for analyzing many such phenomena—I have mentioned conscious perception, rule-following, free will—in physically unexceptional terms and for believing that if the physical terms on which the world operates are fixed, then the psychological will have been fixed as well. But I do admit that there are many sneaking intuitions that suggest it cannot be so. My response is to suggest that those backsliding intuitions simply come of the fact that in the cases in question we lack the capacity to habituate ourselves in the connections alleged and to recognize their effects. Habituation in these cases—habituation in deriving recursively representational states of any kind—can develop only by virtue of being able to experiment with the effects on the represented-ways-things-are of variations in the physical conditions of the brain. And as science and technology currently lie, that sort of tutoring is denied us.

We can all sympathize with the person who knows that there is the gestalt of a cube present in a pointilliste drawing but who just cannot get it to pop out perceptually or, worse, who does not even know what pop-out is like. The lesson of this essay is that if sympathy is due in that case, it is also due on a very much wider front. Those of us who believe in the a priori derivability of the psychological from the physical know, or think we know, that all psychological profiles are present in purely physicalistic configurations. But we are condemned to a position in which we just cannot make that fact salient and irresistible for recursively representational states. We are deeply deprived, at least for the moment, of any sense of inferential pop-out.

Acknowledgments

My thanks for comments on earlier drafts to Colin Klein, Victoria McGeer, Susanna Siegel and, in particular, Jessica Wilson; she wrote an excellent commentary for the presentation of an early version to the Kline Seminar at the University of Missouri, Columbia, November 2004. I am also grateful for the many comments received from the participants in that seminar and for useful conversations with David Chalmers and Daniel Stoljar. The line taken in the essay can be seen as a special case of the more general argument in Stoljar 2006b. At the final stage of preparing the paper, I benefited enormously from discussion at a conference on scientific reduction at Tilburg University, and I must also record my gratitude for some very helpful comments that I received on that occasion.
Notes

1. The problem is related in a variety of ways to the problem of the explanatory gap that Joseph Levine (1993) made famous. I look at the deficiency that he identified with a view to derivation in particular, rather than to explanation more generally, and I attempt to identify sources of the difficulty.

2. For all that decision theory supposes, indeed, the subject need not even be aware of the prospect as attractive; it may be attracted without having any belief to the effect that the prospect is attractive or attracting.
Whatever philosophers say they are doing, one thing that many philosophers seem to be doing is “analysis” from the armchair. They put obvious truths about a subject matter together and see what results, or they engage in thought experiments to try to make their view plausible or a rival implausible; they draw out deductive (or near-deductive) consequences of their own view or their opponent’s view, and take themselves to have established something important: all in all, much of this does not seem to involve laboratories or surveys.

Furthermore, these processes go on not just when philosophers are trying to work out the meanings of expressions, or the applicability of concepts, but also when they are trying to solve problems that are apparently about the world, rather than our descriptions of the world. Philosophers interested in the nature of rational desire, action, weakness of the will, causation, properties, laws of nature, epistemic justification, or knowledge will sometimes engage in activity that looks more like analysis of concepts than empirical psychology, experimental physics, or empirical sociology.

A group of theories collectively (and originally derisively) labeled the “Canberra Plan” have offered a story about what is going on here. According to this story, what is going on is indeed conceptual analysis, though of a slightly different sort from some traditional conceptions of that activity, and the conceptual analysis helps answer questions apparently about things rather than about the meanings of particular words or the application of particular concepts by solving “location problems.”

In outline, the process is as follows. We select something we would like a philosophical analysis of: causation, color, free will, beliefs, moral value, or whatever. Then we collect together the “platitudes” concerning our subject matter. I am not sure everyone means exactly the same thing by this technical term (see below), but as a first pass platitudes are claims
about the subject matter that reflect our ordinary use of the term (or specialists’ ordinary use of the term, if we are after a philosophical analysis of a term of art, like ‘gene’ or ‘electron’ or ‘social class’). Once we have these platitudes assembled, we conjoin them, and use the resulting information to define a “theoretical role” for the thing we are interested in. Colors are the properties of surfaces of objects that cause certain kinds of experiences through producing changes in our eyes in suitable illumination, and so on, or whatever it might be. It is often useful to represent this theoretical role with a “Ramsey sentence”: an existentially quantified sentence capturing all, or most, of the information in the platitudes. (The Ramsey sentence for the property of being a free action may start, e.g., with A thing such that it is the unique property that is sometimes instantiated by human actions and does not obtain when the agent has no influence over the action, and . . . .) Ramsey sentences are particularly useful when a cluster of things are to be analyzed at once—famously, an analysis of beliefs may need to mention other kinds of mental states (e.g., how they typically go together with desires to produce behavior)—but we can make a statement of the roles of a range of mental states at once in nonmental vocabulary, if we replace the name of each mental property with a different variable. For example, if we were happy to take the notions of ‘action’ and ‘content’ for granted, then we might be happy with ‘There is one kind of thing, and another kind of thing, such that things of the first sort tend to cause rational behavior together with things of the second sort, and both kinds of things can have propositional content, and . . . .’, as a Ramsey sentence to simultaneously give us the roles of beliefs and desires.

The details of regimentation are not so important for the moment. (A canonical source for how to regiment is Lewis 1970.) Equipped with our specification of “theoretical roles” drawn from the platitudes, however we represent those roles, we are ready to take the second step in answering our demand for philosophical analysis. At the second stage, we look at our theory of the world to tell us what, if anything, plays the role so defined. Often the theory of the world we look to is formulated in a more restricted vocabulary than the theory with the vocabulary we initially used in saying what we wanted to analyze, though the theory of the world will hopefully have room for the vocabulary in our theoretical-role specification. We might use the deliverances of color-science and psychology to see what deservers are available for our “color” theoretical role, for example, or (one day!) use our theories of the working of the brain to look for deservers for the causal roles specified by our theoretical-role statement for beliefs, desires, and the rest.
Often the fit may not be exact: perhaps none of the structures in organic chemistry exactly plays the role that “genes” are supposed to, or perhaps the brain does not quite have any kinds that behave exactly as we pretheoretically suppose beliefs and desires do. Then the philosopher may have some more work to do in working out which of the available options is the “best deserver” to be the phenomenon under discussion—which property or thing satisfies the most of the role specified, or the most important aspects of the role satisfied. (The philosopher may also need to decide what to do if there are multiple equally good deservers, or when the best is still pretty terrible.) Some principles about “weighting” different platitudes will probably be employed, though this step of adjudication remains more of an art than a science. Once a best deserver has been identified (or the appropriate excuses made), the philosopher’s job is complete: we will have an account of what color, or causation, or free will, or genes are, and ideally an account in some more perspicuous terms than the ones in which we started. (Partial success is of course possible too: we may have a conclusion about what general sort of thing our phenomenon is, or we may be able to rule out some alternative specific proposals even if a positive proposal for analysis remains elusive.)

Notice, however, that this “Canberra two-step” only helps us to explain how this conceptual analysis could be useful when we already have the answers required for the second step—when we already know what the world is like that we are going to place our phenomena in. So this will be of little help when we come to do fundamental metaphysics: when the Humean about causation and laws disagrees with a non-Humean about causation and laws, it is not a dispute about which of the commonly agreed-upon entities best fills a pair of theoretical roles. Someone, for example, who believes there is a primitive relation of singular causation believes in something in the world that the Humean does not. Likewise when we come to the debate between the dualist and the physicalist about the mind: it is not that the physicalist typically believes there is a complex of nonphysical states with primitive intentionality associated with our brains, but is merely wondering whether our minds are those states or some physical states of our brain (perhaps identical to brain states, perhaps higher-order functionally defined states). No—the dualist believes in a range of states that the physicalist does not believe in, and it is not very clear how arguing over the concept of ‘belief’ is going to settle this issue.

So there are some metaphysical questions that will not be susceptible to “location” solutions, without some further work to sort out what there is in the world for the phenomenon in question to be identified with. There
will be many more questions where it will be to some extent controversial whether they are mere “location” questions. Many people treat debates about personal identity over time as location debates—there are sequences of minds (or stages of minds) that are psychologically continuous, sequences of bodies that are physically continuous, various patterns of informational and causal dependence, attitudes and behavior of groups of bodies and psychologies. . . . Which, if any, of these sequences defined in these terms are sequences of the same person? But others think we need some additional metaphysical resources if we are to identify the best candidates to be the relation of identity over time in this world. Perhaps we need haecceities or essences, connected in some way to the distribution of qualities at times so as to yield distinguished sequences. Perhaps we need a new relation of nonreductive constitution that people can stand in to bodies and psychologies. Maybe we need changeless souls to preserve us in a Heraclitean world of change and decay; maybe we need irreducible substance-sorts . . . or whatever. Can we derail the location program whenever we have a particular metaphysical problem to solve just by making some sort of prima facie case that we need a new fundamental metaphysical commitment? If so, then conceptual analysis will be of little use in most of metaphysics, except perhaps to resolve debates between people who are sufficiently committed to a particular ontology that they will not revise it because of the metaphysical issue in question.

This is a particular worry if solving location problems (and doing the philosophy of language necessary to get the correct machinery in place) is all that philosophy can do, or should do. If the second part of the two-step is something that we should leave to scientists, for example, then it does not look like there is much left for philosophers to do in metaphysics. This would be particularly bad news, given how bad many scientists are at metaphysics—they often don’t know how to do it, and don’t care. Which is sometimes fair enough—not everyone can do everything—but it does suggest there’s a range of important issues that won’t get properly addressed by anybody. Presumably what we should think is that philosophers should not limit themselves to just being in the business of solving location problems—this is a conception of philosophy that could eventually be as stifling as the Wittgensteinian conception of philosophy as only diagnosing and dissolving confusions resulting from misunderstanding language, confusions that according to Wittgenstein and his followers were often due to philosophers in the first place.

Fortunately for the Canberra Plan, and arguably fortunately for metaphysics, the project of assembling “platitudes,” and seeing what they
jointly say about the subject in question, can be defended as valuable even when there is an issue of the second kind to be resolved. This essay explores three ways this method can assist in the second project. I do not know whether all of these usages should count as “conceptual analysis”—perhaps we will need to solve the “location problem” for the concept of conceptual analysis itself to tell! What I do hope to show, though, is that much of the method is of a kind with the method that Jackson, Lewis, and others take to be what we are doing when we are doing conceptual analysis proper. In doing so, I will have pointed out more options for the method, and helped to solve the puzzle I began with: that is, the problem of what philosophers could justifiably be doing when they employ armchair methods to solve metaphysical problems. The significance of this goes beyond metaphysics, in fact—there are many areas of theoretical inquiry where something like the methods of conceptual analysis seem to be at play, for example where people engage in thought experiments. Often what people are aiming for is not just information about our descriptive capacities or our concepts, but rather information about the world. (Thought experiments in physics, for example, are employed in an inquiry about what the physical world is like, not an inquiry primarily about how we represent it.) A more satisfying account of how conceptual analysis could help with such questions would be welcome.

Before I outline the three ways I think a platitudes analysis can be helpful, I think it is worth discussing a few issues about what the platitudes analysis itself involves—or at least what it involves for my purposes. Hopefully this will not only serve to clarify what I have in mind, but to sidestep, or address, common objections raised to the Canberra Plan of conceptual analysis.

1 The Platitudes Analysis: Filling in the Details

‘Platitudes’ is an expression that is to some extent a placeholder—it is a name for those claims or principles we use as the raw material for the process I outlined. But before I go on, I should say something more about what we might think these platitudes are. This is not so much because I want to defend a particular answer—indeed, there may be no unique answer, especially if the platitudes analysis might be employed for different purposes on different occasions—but for two other reasons. One is to give a sense of the range of different options there seem to be here, since I have come across different conceptions being employed by different people, and I think there may be some risk of people talking past each other on
occasion. The other is somewhat related. Sometimes objections are made to the method of platitudes analysis on the basis of certain specific conceptions of what the platitudes are and how we are to tell what they are; and although those objections may be good as far as they go, if it is true that the structure of the account can be preserved with any of several conceptions of the platitudes plugged into it, then we should be careful not to reject the framework altogether simply because of concerns about one way of fleshing it out.

So, what are the platitudes? Or rather, a better question might be what the platitudes are for a certain term or concept: for example, when we are analyzing ‘belief’ (either the word or a mental analogue), what sentences or other representations about belief ought to go in our list of platitudes? (Or perhaps when we are analyzing beliefs what we should ask for are all the platitudes concerning mental terms—not just beliefs, but desires, intentions, actions, perceptions, etc.) After all, what we need for any given occasion we want to find out the “platitudes about X” (or the platitudes concerning X, or whatever)—we rarely, if ever, need to know what the platitudes are simpliciter. One version of this question is the question of what kind of thing a platitude is: is it a belief, or a sentence, or a proposition, or something else? If we are analyzing a word or expression, such as ‘belief’ or ‘free will’ or ‘law of nature’, then I suppose it is natural to think that platitudes concerning these will be some of the sentences involving these expressions (or perhaps something expressed by these sentences—if we think the platitudes about laws of nature should be the same in English and French, for example). If we are analyzing concepts, like the concept of belief, or the concept of a law of nature, or whatever, then the platitudes might be beliefs. Or in either case we might say they are propositions—but if they are propositions, they will presumably be structured propositions, given how they are manipulated (see below). If we are talking about analyzing some phenomenon, on the other hand: color or causation as opposed to ‘color’ or ‘causation’, then presumably we might have in mind any of the three tasks best served by looking at interpreted sentences, beliefs, or structured propositions respectively. Rather than being careful about this, let me talk about platitudes as if they are sentences, and when I talk about something figuring in the platitudes, let me be taken to be talking about words or expressions. (So were I to say, e.g., that assertion appears in some of the platitudes about truth, the initial gloss should be that ‘assertion’ appears in some of the sentences including ‘truth’ that count as platitudes for truth.)

One story, perhaps one of the earliest ones, says that the platitudes about some X are all the sentences talking about X that we take to be true. (X
may be a single thing, or it may be a family of things—the platitudes about the mental, for example, presumably include platitudes that include words like ‘belief’, ‘desire’, ‘thinks’, ‘is conscious of’, etc.) Something like this seems to be what is going on in Lewis’s seminal “How to Define Theoretical Terms” (Lewis 1970), where the Ramsey sentence involved in defining a set of terms we want to analyze (the T-terms) is a “T-postulate” that contains all of what our theory tells us using the T-terms. A related story, which is closely aligned with the Lewis (1970) one, is that we identify some theory containing the X vocabulary, and it is only the sentences that make up that theory which are the platitudes for the Xs. Which theory? Here we have a range of options. For technical terms with a limited community of users, we may be able to point to some defining theory. For other terms, we may be able to locate a “folk” theory—as some people hope to find a folk psychology, or a folk physics, or a folk biological taxonomy. (Lewis ([1972 (in 1999), 257] says that mental terms are defined by “commonsense psychology.”) The theory we use to fix the platitudes may or may not be one we currently accept—for instance we could take ‘tree’ to be implicitly defined by a folk biology which we take to be mistaken in places, or if we want to know what the platitudes are about caloric fluid we may take some theory or theories of caloric fluid to provide the platitudes, even though we now reject such theories.³

At the other extreme, we might be very restrictive about which of the sentences we accept should count as platitudes. For example, we might take a platitude about Xs as something someone believes on pain of being linguistically incompetent with the application of the X terms. (Or we may require that someone believes the relevant platitudes as a necessary condition for possessing the X-concepts, to formulate the equivalent constraint on the mentalistic conception of platitudes.) Formulated this way, platitudes are treated rather like traditional analytic/conceptual truths were supposed to be: someone’s competence with ‘bachelor’ could be called into question if they didn’t believe that all bachelors were unmarried.

Formulated this way, traditional suspicion of analytic/conceptual truths translates into suspicion that there are any such platitudes. We might doubt that there is any error that must be attributable to a fault in linguistic competence rather than a mistake of fact, or we might think that very little is required in the way of belief for someone to be linguistically competent with an expression. Or we may think that some expressions have some such constraints, but many do not. (What do we have to believe in order to competently classify something as falling in the extension of ‘cat’?)
Many people are not explicit about what they count as a platitude. But I have a sense that there is clustering at two places on the spectrum between counting every sentence containing the vocabulary that we accept, at the one end, and accepting only the nonnegotiable, at the other. One place is close to the nonnegotiable analytic truths end. Platitudes are truly central to what it is to be an X. (Call this the “Strict conception.”) There are two, or possibly three, respects in which people may vary from a fairly traditional insistence on analytic truths. The first is that platitudes are not quite nonnegotiable: there may be conflicts between platitudes, and in this case (and maybe only in this case), we need not insist that a phenomenon satisfy all the X-platitudes in order to count as an X. One example of this might be the naive theory of truth—an unrestricted T-scheme together with a straightforward naive theory of quotation enables us to prove a Liar sentence is both true and not true: most people think this shows one of these platitudes about truth and quotation should be not accepted as is. The second is that we may wish to treat some sentences as X-platitudes, not because competent users accept them, but because competent users accept that they would be true were there Xs (or if X was instantiated if it is a property, or obtained if it is a state, or whatever). It may be a platitude about caloric fluid that it is transferred if and only if there is heat transfer, or a platitude about God that she or he created the universe. But we may insist, not that competent users accept these claims (assuming they are genuine platitudes), but only that they recognize that they would be true if there was such a thing as caloric, or if God existed. A third variation can be in the kinds of claims one takes to be platitudes. Influenced by Kripke, Putnam, and Burge (Kripke 1980; Putnam 1975; Burge 1979), this view holds that platitudes for Xs of the form ‘being a causal origin of our use of the expression “X”’, or ‘being what the experts intend to refer to with their use of “X”’, or ‘is a natural kind that actually thus-and-sos’ are the sorts of platitudes that we may find on our list for some terms. Whether this variation is a departure from the traditional conception of analytic truths, or merely a variation within the bounds of that tradition, is not clear.

On this strict conception, platitudes define our subject matter in a straightforward way—to disagree with a platitude is to make a conceptual error or alternatively to change the subject. This gives us relatively little flexibility, and the platitudes, if they are to be this central, may not be informative enough to give us a unique fix on our subject matter. Of course, some of those who take platitudes to be this unyielding will be happy to indulge in changing the subject, providing the subject is not
changed too far: a surprising example is Frank Jackson (1998a, 45), who claims that our concept of identity of persons over time requires an irreducible further fact, beyond physical or psychological continuity or similarity, and in fact requires something that is “never instantiated.” Jackson is committed to the view that there are no people who exist at more than one time, given the meaning ‘people who exist at more than one time’ has. (Since I doubt physical systems like us could be instantaneous people, I presume the thing to say on Jackson’s view is that there are no people. I wonder whether there is mentality and agency or any of those things either, given what Jackson requires for there to be people?) Unsurprisingly, Jackson advocates that we use a replacement concept—a “different and ‘nearby’ conception,” to use his words. In the language of platitudes, Jackson believes there are platitudes about personal identity (or about personhood and time, perhaps) that a relation of identity over time for persons must satisfy—and these platitudes are sufficiently nonnegotiable that a relation that does not satisfy them is not identity through time for persons. Occasionally I have heard talk of “changing” or “adjusting” or “revising” the platitudes associated with a given term—I am not sure whether this is talk about choosing to use a different expression with a different sense (but which is spelled and pronounced the same), or whether it presupposes that we can change the meaning of an expression in our language just by making a decision about it (perhaps this is true for some technical terms, but I doubt philosophers usually have sufficient linguistic authority to change our language or our conceptual scheme in this way), or whether perhaps this is talk of something else.6 No doubt the speakers are not always clear what they have in mind either.

There is an alternative conception—I don’t have a group to pin it on yet (maybe it is idiosyncratic). This alternative conception falls closer to the “kitchen sink” end. Not everything we believe (or that is claimed by a defining theory) need go into the mix, but quite a bit does. Going along with this, platitudes are conceived of as being rather more fallible, and individually the X-platitudes are less like necessary conditions to be an X. The place where this kind of conception is most in evidence is in analytic functionalist treatments of folk psychology. Many of the psychological generalizations we are inclined to believe will, no doubt, be shown to have counterexamples, or not to apply straightforwardly to unusual cases.

The progress of cognitive psychology should show us that there are things still to be learned about the behavior of beliefs, desires, intentions, emotions, and the rest. Not every generalization we believe about ordinary psychology will be vindicated, and it may be fruitless to try to identify any
particular generalization that we expect will be unrevised. But according to analytic functionalism it is still the case that whatever beliefs, desires, and the like are, they satisfy many or most of the generalizations about their behavior provided by folk psychology. Nevertheless, if the only “mental” states mature psychology were to deliver us satisfied completely different generalizations and claims about their interrelations than the folk-psychological theory about mental states, then that would show that there were no such states as the familiar beliefs, desires, hopes, fears, pains, emotions, and all the rest.

Once a much more generous conception of platitudes is embraced, then we need to say what the limits on this generosity are. If we allow every claim we believe about a given topic to be a platitude, then we are back with something like the Lewis (1970) view. (Lewis was talking about theories rather than sets of beliefs, but we can treat a person’s beliefs as a special case of a theory, albeit an often inarticulate and messy one.) Even at this extreme, there are still different ways to go. When I am considering what the platitudes are about personal identity or free will, I might include all of what I believe about the topic (or at least as much of that as I have straightforward access to.) Or maybe I should impose a requirement of common belief—perhaps, even at its most generous, the platitudes are only the things that are believed in common in my community. Or perhaps we should be even more restricted, and include only the things that are not only believed throughout our community, but about which it is common knowledge that they are so believed.

If we are to be less individualistic, and decide on what are platitudes on the basis of shared beliefs, or shared beliefs of which it is common knowledge that they are shared, then the question also arises at this point about which community is the one that we should be concerned with. All of humanity, or all of contemporary humanity? People with a similar culture, or similar patterns of engagement with the world? People who think as I do (which may mean, in effect, that what we believe is just what I believe)? I expect our verdicts here may vary from case to case. In the case of platitudes employing technical terms, or terms that began as technical terms, such as ‘electron’ or ‘supervenience’, or ‘gene’ or ‘entropy’, we may be primarily interested in the beliefs of experts, somehow identified. In the case of “folk” notions, like ‘thinking’ or ‘heat’ or ‘artifact’, we may think the community needs to be selected more inclusively.

One limitation on the size of the relevant community may be linguistic. I have been talking loosely of platitudes being given by what is believed, but strictly speaking platitudes are going to be sentences (interpreted
sentences). Given this, we may wish to restrict our communities to those sharing a common language. Those who speak no English are unlikely to accept the sentence ‘Some beliefs are stronger than others’—or any other English sentence, for that matter. Of course there is linguistic variation among English speakers as well—we will get a very strange picture of what it is to ‘luck out’, for example, if we do not restrict our platitude gathering to one or other of the patterns of use of that expression. (North Americans use ‘luck out’ to describe situations in which one is lucky, whereas British and Antipodean speakers use it to describe situations where one is out of luck.)

Going to what is held in common in a linguistic community is one way to narrow down the platitudes to something less generous than just including any old thing I happen to believe on the topic. Another way is to include among the platitudes only the things that I take to be clearly true, or obvious, or certain about the relevant topic. When considering the platitudes about possible worlds, for example, the claim that a world has some things true according to it is plausibly platitudinous: that there is a world that contains more individuals than sets is not platitudinous, even though it is something I am tempted to believe. We could explain this by saying that the second belief is idiosyncratic (which would invoke some standard of commonly shared belief), or we could explain this by saying that the second belief is tentative, not something I take to be obvious, and something that I find myself inclined to think requires justification in terms of more evident claims about modality, possible worlds, and so on.

If we raise the standard for something to count as a platitude in this way, then platitudes become more like traditional philosophers’ “intuitions”—though notice that clearly true or obvious claims need not be particularly a priori or philosophical. It is a platitude of human physiology that one’s nose is typically smaller than one’s foot (albeit that it is platitudinous in the ordinary as well as the technical sense): but we need not suppose that this claim about noses and feet is distinctively philosophical, nor discoverable through some privileged faculty, nor a conceptual truth, nor any of the other things that some have been tempted to think about typical philosophical “intuitions.”

Another constraint we might try to put on platitudes involves their origin. There are some things that one has a tendency to believe for specific reasons, and other things that one is inclined to take for granted. This can vary from context to context, of course, and it is also true that what is taken for granted by somebody may be quite specialized, technical, and/or
esoteric information: what a particle physicist takes for granted about quarks will not be much like an unquestioned dogma of everyday life. Perhaps when people are assembling platitudes about the mind, or causation, or moral rightness, or whatever else is the target of their philosophical investigation, some sensitivity to whether a claim is taken for granted or seems eminently contestable may play a role in the sorting procedure.

Once we have criteria that come in degrees, such as degree of common belief or common knowledge, or degree of certainty, or degree to which something is taken for granted, we are provided with a way of assigning “weightings” to different platitudes, in proportion to the relevant quantity. Exactly how to weight platitudes is a vexed question, especially if more than one factor that comes in degrees makes for platitude status. Questions about how platitudes are weighted are of particular importance when platitudes conflict, since people often think that one platitude can trump another (or one group of platitudes can trump another group) if one is much more important or weighted much more highly than the other. I have the impression that those engaged in the Canberra Plan do make judgments about relative weightings of platitudes, and these make a difference in decisions about what to reject when there is a conflict (or when nothing in the world precisely satisfies the original list of platitudes), but I am less confident about what explains these judgments.

In making the above remarks, in part I am trying to characterize what is going on when those who are generous about ‘platitudes’ sort platitudes from the totality of what they believe about a topic, or the totality of what a theory they accept says about a topic. The characterization has not been altogether precise, in part to give a feel for a range of variation, and in part because I am not entirely clear myself what unarticulated principles are being deployed. But one of the things I mean to be doing in providing the above sketch is just to indicate an end-point on a continuum of positions. At one end is the strict conception of platitudes as more or less traditional analytic truths, and at the other end are those who include most of what they take to be true, or take to be common belief, or that is pretty clear or certain or obvious about a given topic. Individual people who find it useful to assemble platitudes as part of philosophical investigation may fall closer to one end or the other of this spectrum (or may even fall in different places on it on different occasions, or when dealing with different subjects). Or I suppose for many people who engage in this project it might be difficult to tell where they fall on this spectrum, or there may be no very determinate fact of the matter.
2 What to Do with the Platitudes

Whatever the account of what platitudes are and how we collect them, eventually we come to have a collection of sentences that we take to express platitudes about a particular subject matter (let us assume for convenience these are all in a single language). We can then use this collection to discover things that might not have been obvious from our implicit grasp of the platitudes taken in isolation. For a start, it may only be after this process that we come to see things that require entertaining several platitudes at once. One example is illustrated by an argument for compatibilism in the free will debate. Initially, we may have the intuition that our actions being determined by past events beyond our control is inimical to free will. We may also have the intuition, though, that if our actions were random that would also not be a case of free will. When we put those together with the suggestion that there isn’t much more (or maybe anything more) to something’s being random than its not being determined by any previous state of the world, we might be inclined to look more favorably on the idea that some sorts of determination might be compatible with freely willed action. Each principle (the one about determinism precluding free will, the one about randomness precluding free will, and the one about randomness following from a lack of determination) has some initial plausibility—each has some call to be called a platitude, at least on suitably generous conceptions of platitudes. But it is only when we put all three together, or put all three together with the “platitude” that some of our actions are freely willed, that we see that there might be good reason to revise our principle about the incompatibility of free will and determinism. Of course, whether the adoption of compatibilism is the right response to the discovery of this tension is controversial. But I am merely using this as an example of how one might think some important philosophical insight may come from considering together thoughts that are individually platitudes, or at least seem to have some claim to that status (especially if we have a relatively generous conception of platitudes). Note that this example is a case where the compatibilist at least can be seen as arguing that one of the platitudes is, in the end, incorrect—so it sits less comfortably with the picture of platitudes as core conceptual truths about a subject matter.

In the account of solving “location problems” I gave in the opening section, I more or less took it as read that we would have a theory available to tell us what the deservers for the platitudes are. But there should be more to say about how this part of the task is carried out. Sometimes it
will be controversial what the world is like. Sometimes the only theories of the relevant bit of the world we will have available will employ the vocabulary or concepts under investigation. And sometimes, perhaps often, our theory of how the world is will not deliver up any things that precisely play the family of roles given to us by a group of platitudes. I expect people employing platitudes analysis can fruitfully disagree about what to do in these situations.

There is one final dimension of possible difference in approaches to platitudes analysis that I want to mention. This range of differences centers on our justification for thinking that the “deservers” identified by this method really are the referents of the terms or concepts that appear in our network of platitudes. A view traditionally associated with the platitudes analysis is that the platitudes associated with a group of expressions together give us the meaning of those expressions, and so it is analytic that any group of things that jointly satisfy (or best satisfy) those platitudes are the referents of the given expressions. Many of the proponents of a platitude approach are descriptivists about the meanings of theoretical expressions, at least: Carnap developed the method of employing Carnap sentences to capture the meanings of theoretical terms in part as a response to Quine’s attack on analyticity (Carnap 1963, 958–966), so that the platitudes analysis was capturing the meaning was important to him. Lewis suggests that this method gives us analyses of the relevant expressions, where this seems to be a claim that it serves up analytic truths (see, e.g., Lewis 1989; see also Nolan 2005, 213–227). And Jackson is a well-known defender of descriptivism about the meanings of a range of expressions (see, e.g., Jackson 1998a), and his defense of analytic functionalism about the mind makes clear that the folk-psychological platitudes are supposed to deliver the meanings of mental vocabulary, as well as enable us to determine the extensions of the relevant predicates and referents of the relevant property names (Braddon-Mitchell and Jackson 1996, ix, 46).

Descriptivism about theoretical terms may have much to recommend it. But note that for several of the core purposes that these sorts of analyses might be used for, it is not required. Suppose you thought that names of properties associated with natural kind terms were directly referential. For instance, let us suppose the expression ‘Being a whale’ took as its semantic value simply the property of being a whale. You assemble the “platitudes” about whales: these might be the platitudes among experts, if you are particularly interested in the expert usage, or they might be platitudes in the community at large. Suppose, in any case, the platitudes you assemble
make claims that you at least believe are true, or true for the most part. (Perhaps you defer to whale experts and believe what they tell you about whales.) Suppose, when you put the platitudes together in the appropriate way, they tell you that being a whale is the unique property such that it Φ, and its instances behave thus and so (swim in the ocean, are such that some of the instances eat plankton, are all such that they give birth to their young live, etc.). You might then use this information to go looking for the property such that it Φs and its instances behave thus and so. Suppose you find one, describable in alternative vocabulary. (Let us suppose that the property of being an animal with DNA of such-and-such composition is, indeed, the unique property such that it Φs and its instances are thus and so.) It seems that you are in a position, given what you know, to infer that being a whale is identical to the property being an animal with DNA of such-and-such arrangement. And it seems that you are able to do this without making any assumptions about descriptivism versus direct reference for the expression ‘being a whale’. If you believe the platitudes, know that they imply that being a whale is the unique property such that it Φs and its instances behave thus and so; and know that being an animal with certain DNA is a unique property such that..., then that is all you need for the inference.

At least when the platitudes employed are ones you believe, or when you believe that they are largely correct, you will have reason to think that if they have a good enough group of things satisfying them, then those things are the things referred to. This is so even if you have no idea how the expressions get their referents. You might be torn between descriptivism and some sort of direct reference (or some third option) for mental vocabulary, perhaps even because you are not sure whether the expressions behave more like natural kind terms or like functional vocabulary, because you are not sure whether, for example, beliefs form a natural kind or only share a functional specification. You can use the platitudes about mental states as a guide to what the property being a belief amounts to, and then see what sort of property (functional, natural kind, other), if any, does well enough satisfying that specification, and then hypothesize about how that property is associated with the word. As Carnap pointed out, it is your theory that commits you to the corresponding Ramsey and Carnap sentences, not anything else. The further claim that the associated Ramsey or Carnap sentences have much to do with the meaning of the terms in your theory is an extra step, of course, but producing a specification of roles and hunting for deservers does not need this step, if all one is interested in is what in fact the terms refer to, if anything.
So taking the network of platitudes to capture some or all of the meanings of the relevant expressions is an optional add-on, it seems to me. It does bring some benefits, of course. One advantage is that it would enable us to return verdicts about merely possible cases as well as actual cases. If some hypothesis is entailed by ordinary analytic truths, it will be necessary. So if we are interested, for example, not just in the question of what beliefs and desires happen to be, but also what the limits are on what they could possibly be, then if we can take a group of insights into beliefs and desires to be analytic we will have much more guidance on the broader question. A related advantage is that if the conclusions one reaches are supposed to be analytic, then they can be tested more easily by considering merely possible cases. If the claim that people often desire to do what they believe is right is merely a commonplace, then it may tell us very little about whether there are possible creatures that do not have any desire to do what they believe is right. But if it is a putative analytic truth about ‘right’ (or the combination of ‘right’, ‘belief’, ‘desire’, etc.), then we can test to see whether it has that status by considering the intelligibility of a thought experiment where we describe such creatures. Finally, if we both thought that platitudes were analytic, and that analytic truths were implicitly grasped by those competent with the relevant concepts, then we might be more confident that purely armchair methods will enable us to discover them. If I am confident that I competently deploy the concept of causation (and competently use the English word ‘cause’), I can be confident that I already know the platitudes about causation, at least implicitly, and all I need to do is whatever self-examination is required to make them explicit. If they are not analytic, I need some other reason to be confident that I can come up with them merely through reflection. Likewise for platitudes about truth, moral rightness, color, personal identity, and so on. If I had good reason to believe both that the platitudes were analytic and that the analyticities were all things I implicitly knew, I could have methodological confidence in armchair philosophy that might otherwise be harder to come across.

Despite this appeal, some will not want to take the extra step of identifying the meaning this way. They might, for example, be convinced that this sort of descriptivism about theoretical terms is ultimately mistaken. Perhaps they are convinced that the world can be sufficiently surprising so that competent language users cannot tell in advance that many of the platitudes are correct (maybe science could have determined that electrons are a kind of German sausage after all, if the evidence had been sufficiently bizarre?), or perhaps they subscribe to a philosophy of language according
to which meanings are determined by word–world connections that provide for very little analytic truth. Or they may even be undecided, or wish to stay neutral—they may have enough challenges of their own to worry about, without taking on a controversial commitment to descriptivism. The benefits for metaphysical inquiry I will be suggesting should appeal even to those who reject, or are suspicious of, any straightforward connection between platitudes and meaning.

3 How the Platitudes Can Help: The First Way

The difficulty, it seems, is that when we are doing fundamental metaphysics we are trying to work out what the world is really like, but all that the method of putting together the platitudes gives us, according to Jackson (et al.) is something else—it gives us information about a term or concept that, when put together with a story about how the world really is, tells us what that term or concept applies to. The first way I want to mention in which putting the platitudes together can help provide answers to metaphysical questions is a relatively familiar sort. Getting clear on our concepts and meanings can help us disentangle different options, it can give us information about what is conceptually possible, and it can help sharpen the questions we then attempt to answer.

These are traditional tasks for conceptual analysis, and it may explain some of the armchair use of platitudes in metaphysics—especially the use of deductive reasoning to show things that are not obvious to competent users of the concepts. If it could be shown that omnipotence and omnibenevolence were mutually inconsistent, we would not need to investigate the world to see if there was an “omni-god” as conceived by some strands of traditional theism. If the occurrence of change was inconsistent with there being a determinate future, this might drastically constrain our options for a metaphysics of time. If we have to suppose that mental states sometimes are caused by perception and cause behavior, we can rule out epiphenomenalism about mental states (at least across-the-board epiphenomenalism), without needing to investigate whether there are any non-physical epiphenomenal states.

In a similar vein, becoming aware of slight variations in meaning and views can be valuable preparation for doing good work, in metaphysics and elsewhere, and it is something that philosophers in the analytic tradition typically pride themselves on. Whether it is pointing out that the questions of what a cause is and what is relevant to a causal explanation are different questions; or that the hypothesis ‘everyone always acts
selfishly’ has a range of different meanings ranging from the implausible to the possibly quotidian; or that ‘truth’ for sentences and ‘truth’ for propositions would be quite different, though presumably connected—distinguishing options is something that can to some extent be done usefully in the armchair, on the basis of little more than a grasp of the concepts or the meanings of the relevant terms.

Another thing that collecting the platitudes can do is help define our subject matter (see Jackson 1998a, 30–31). If you ask someone off the street whether there are any properties, and if so what are they, my guess is that person will have little idea what you are asking (and this is true even after we clear things up a bit—by making it clear we are not talking about real estate, we are not talking about whether people sometimes use abstract nouns, and maybe even after we even give some examples of putative properties like colors and shapes, etc.). One thing that collecting platitudes can help us do is to get a clearer handle on what sort of question we are asking. Would there have to be things that were multiply located in order for there to be something that counted as a property? Or is that just an optional add-on (which, e.g., we may want to include to help answer the question of what it is for a thing to instantiate a property, or we may want to reject when the time comes)? Do properties have to be causes, or effects? Do the properties had by distinct particulars have to be numerically identical, or not? Some of these matters may be able to be settled by old-fashioned conceptual analysis—some claims about properties may be so central that to deny them is tantamount to either contradicting yourself or changing the subject.

Removing conceptual confusions about possible answers, distinguishing possible answers that may otherwise be run together, and getting clear on what are the metaphysical questions are all fruits offered by traditional conceptual analysis. They are the sorts of things that should be valuable across the board in almost any inquiry about the world, and not just metaphysical inquiry. They may be particularly useful for the kinds of questions that philosophers address, of course, because those questions can be very general, and lack a generally agreed methodology to answer them. (And other disciplines often have more clearly defined questions to address, or so it seems to me—though I suspect coming up with the right question to ask is important in virtually every field, even fields with better defined methodologies for answering them.)

I suspect the platitudes approach will deliver less along these lines than traditional conceptual analysis was supposed to. This is because the platitudes approach is likely to yield a cluster of descriptions, any single one
of which is likely to be negotiable if a candidate does well enough in satisfying the others. (So we may need to revise our offhand opinion about how many objects there have been in the room in the last week, if we accept the existence of temporal parts or subatomic objects, and we might need to accept that things come to be at different temperatures when the same amount of heat is applied to them, or to give up other commonsense views.) We might have to live with a few thought experiments delivering counterintuitive results, or a few commonplaces being overturned. This can be seen either as bullet-biting, or as making exciting metaphysical discoveries, according to taste. This makes things more messy—bullet-biting can seem ad hoc, and questions like the relative importance of different platitudes are difficult to settle and potentially endlessly controversial—especially given the human tendency for proponents of a theory to weight highly the respects in which a theory does well, and minimize or disregard the respects in which a theory does poorly.

On the other hand, we do not get to choose whether or not our methods are going to need to be messy. I would like a sharp, clean method in conceptual analysis as much as the next person—a way of showing that a counterexample is indeed fatal to a view, or can in good conscience be resisted provided some definite set of criteria is met. Perhaps one day things will be like that—Leibniz’s dream that philosophical disputes could be settled by mutually agreed-on precise algorithms might be down the track somewhere—but I’m not holding my breath, and I don’t recommend anyone else does either. I am convinced by some of the fairly standard examples that definitions can be widely accepted and core platitudes can be taken to be extremely important, and yet turn out to be mistaken. To take an example from the history of physics, everyone accepted that the definition of momentum was that it was a quantity equivalent to mass times velocity. This wasn’t even taken to just be a well-confirmed fact—it was taken to be a definition. One might argue whether momentum was constant in all interactions, but not, one would have thought, whether it was directly proportional to mass times velocity. Relativistic physics convinced us that momentum was not quite proportional to mass times velocity, however. In fact, the quantity of momentum as relativity conceives of it can be indefinitely different to the product of an object’s mass and velocity (at least its rest mass and velocity). Instead, we think that there was enough about the Newtonians’ use of the word ‘momentum’ that this quantity is what they were talking about—it is conserved in reactions, is roughly proportional to mass times velocity at low speeds, it explains the observations that led us to use the expression ‘momentum’ in physics in
a fairly straightforward way, and so on. The lesson I want to draw here from this case is that we are likely to overestimate the importance of any single platitude (such as the official “definition” of momentum, which was a core platitude). And we should keep in mind that something may turn out to be close enough to the set of constraints that are in place in unexpected ways. On the other hand, although some may use the momentum case to argue that no combination of platitudes constrains what ‘momentum’ may apply to, I do not think it shows this—momentum couldn’t have turned out to be a kind of cat, or be identical to the temperature 24°C. (To be old-fashioned, I’m even prepared to suggest that these claims are analytic—you can be in a position to know that momentum isn’t a kind of cat simply in virtue of understanding expressions like ‘momentum’ and ‘cat’.)

So I am suspicious when someone claims to solve a traditional philosophical problem—the nature of free will, how objects continue in existence through time, the nature of causation—through the platitudes analysis (or other sorts of conceptual analysis) alone. Our terms and concepts probably have too much slack to settle these questions. But this is not to say a platitudes analysis is not helpful in these ways—I think there are useful contributions that conceptual analysis in general, and the platitudes analysis in particular, can make to investigations of the world—and crucially, these are contributions that can be recognized independent of knowing the details of what is in fact the case. Recognizing analytic constraints on what descriptions of the world are coherent is a traditional conception of philosophy, or at least part of it. That conceptual analysis can provide useful distinctions and that it can help us to get clear on our subject matter are not exactly new ideas, either. Together these enable the platitudes account to help play the “underlaborer” role in investigations of the world, something like the role that Locke suggests for his philosophy. This applies to metaphysics as much as to other inquiries about the world—the platitudes analysis can stand to metaphysics as it stands to other inquiries, including many outside the current disciplinary bounds of philosophy.

4 How the Platitudes Can Help: The Second Way

The above virtue of the platitudes analysis is one that we should expect defenses of conceptual analysis more generally to claim for their enterprises. The other two benefits from a platitudes analysis are less like the sorts of advantages that might be expected from traditional conceptual
analysis. This might be an advantage of the platitudes account over some traditional forms of conceptual analysis: it may also be a reason to think of the platitudes analysis as not being a kind of conceptual analysis at all. One feature both of these benefits have is that the support they provide for a view is typically a posteriori—which is not part of the original picture of conceptual analysis!

Assembling platitudes about a subject matter (whether causation, or free will, or genes, or acids, or whatever) often results in providing a statement of folk theory—or at least the less speculative parts of folk theory. For example, if we were trying to assemble the platitudes for ‘tree’ (in the sense of a kind of vegetation), we would mention trees being plants, trees usually having leaves, trees being largely made of wood, and so on. Maybe some part of this story would be analytic, or definitional, but much of it would go beyond that. Articulating widely shared implicit beliefs employing a concept (or stated using a term), together with drawing consequences from the body as a whole, will probably reveal a theory that will be incomplete in all sorts of ways but will likely articulate a set of commitments we may in fact already have, or at least that we are inclined to have. Or when it is the platitudes of some group of specialists we are investigating, we may get a set of commitments that our experts share and we are inclined to defer to. (Or on occasion someone may be both a conceptual investigator and a “specialist” in the relevant sense, as when a theoretical biologist tries to articulate the concept of a ‘gene’ in population biology, or a philosopher tries to do a conceptual analysis of what it is for something to be a proposition or a possible world.)

Assembling platitudes and seeing how they fit together is obviously not the only way to discover what we believe. Often we can tell a lot about what our views are just by introspection, and there are also other ways that require little more than an armchair—I often discover things about what I believe in conversation, or by reading or listening to someone else articulating what they believe, and trying to work out whether I agree. Coming up with my “pretheoretic” opinions about a topic may be done in a similar way, though there is always the problem of contaminating one’s pretheoretic views by paying them too much attention—it is hard to avoid slipping into theorizing, and hard to avoid attributing to one’s past self the beliefs one thinks one ought to have had. If my only task was to come up with what I believe, coming up with platitudes might seem a fairly inefficient way of doing it. But there are several things platitudes can be relevant for in this area, some of which are harder to do by mere introspection.
In metaphysics as elsewhere, we form a community of inquiry. (Less so than in many areas of research, I’ll grant.) As well as working out what I believe, I often want to know what we believe: we might be the philosophical community at large (or some subsection, as when I try to work out what I think the orthodoxy is on a given issue), or it might be some subgroup who is sympathetic with my views about some issue (what do the realists believe, or what do the objectivists in ethics believe?), or it might be narrower again (what consensus is there in my current conversation, or among the people I talk with about a certain question?). Again, asking people or reading what they have written on a topic can often serve very well for answering these questions: but the kind of investigation to find out what there is agreement about will be much the same as investigation into what the platitudes are for the relevant group, especially if we ask about relative certainty and relative unrevisability.

Another thing we might be interested in often goes under the name of ‘common knowledge’: what are the things that are not only believed by a group, but that the group believes each other to believe, and believes each other to be aware of this in turn, and whatever else distinguishes common knowledge from knowledge that happens to be had by each of the members of a group (or most of the members of the group). The only problem with the label ‘common knowledge’ is that the thing we are sometimes concerned with is the consensus—whether or not that consensus is correct, or justified. Something like ‘common belief’ would be a better term for this—so let me use that expression to cover those things that are believed, that a group largely believe each other to believe, and are aware of this mutual belief, and so on. It is more than what is in fact believed by each of the group. (It may also be less, if there can be common knowledge or common belief in a group even if the occasional member does not believe it, or is not aware that it is commonly believed, or whatever.) In determining what the common belief is, or what the public consensus is about a given topic, a platitudes analysis might be the best bet. Indeed, on one possible account of what platitudes are, the platitudes just are what is commonly believed, though facts about how they are to be weighted may go beyond information simply about what is commonly believed in a community.

A platitudes analysis might also be what is required when we want to find out what is believed by a group that we do not belong to: if I was trying to assemble the platitudes about wave-function collapse or shamanism or Russian literature I would not put a great deal of stock in what I believed. When we are trying to discover what the platitudes are among an expert community for some expression or concept, the result may well
be a map of what are matters of widespread, relatively uncontroversial belief among that group. This information can be useful and interesting, even when it goes beyond telling us conceptual truths about the expression or concept in question. I claim that it can even be useful for advancing our knowledge or understanding of the world, and not merely the expression or concept—our understanding of wave-function collapse, shamanism or Russian literature, for example.

The characterization of platitude-collection in the previous few paragraphs goes better with views of platitudes that are toward the “kitchen sink” end of the continuum discussed in the early sections of this essay. Presumably it is harder to read off which platitudes, if any, capture the unrevisable core than to tell what people, even experts, happen to believe, though perhaps we could at least suppose that it is a necessary condition for a claim being part of the unrevisable core of claims employing a concept that all the experts agree it is true. To the extent that part, or all, of assembling the platitudes is discovering a “folk theory” about a topic (or an “expert theory,” I suppose, if I am deferring to experts in a field), my remarks above should apply.

Self-discovery about what I believe is all well and good—but how is it relevant to what is in fact the case? Suppose I do discover that I believe X, Y, and Z about free will, or causation, or time. What sort of help is that in working out what free will, causation, or time really are? (Setting aside the sort of benefits I mentioned in the previous section.) One way it helps is instrumental, or “pragmatic”—it can help to keep track of what you yourself think, while trying to improve your evidence about a given question. People who do not seem to realize exactly what they believe might get more easily muddled, or have a harder time in checking whether belief transitions they make during an investigation are justified. And being able to check one’s belief transitions is important for most of us if we are going to avoid going wrong. Finding out what we believe, or what is ‘common belief’, or what the experts believe can be important as well. To the extent that metaphysics or anything else is a communal inquiry, one thing a lot of us as individuals will be wanting to do is keep an eye on what the community of investigation thinks—the better to critique it, or extend it, or provide arguments to others that proceed from premises they are willing to grant. Having good information about this sort of thing is like having time and funding—it isn’t itself evidence of a view, but it is important to coming up with views that are closer to the truth. (We hope!)

I think it can also make an evidential difference what beliefs I have. Many writers have endorsed the view that we have some reason to
continue to hold the beliefs we currently have, because we currently hold them, rather than change to a set of beliefs otherwise equally supported by the evidence. This is often called “epistemic conservatism.” I should note that when I say it makes an evidential difference, I mean this in the sense that it makes for something epistemically better, ceteris paribus, rather than being of merely pragmatic use. (Or something close to ceteris paribus epistemic betterness, anyway—there is a useful debate to be had about how the evidential theoretical virtues relate to other epistemic evaluations.) Some adopt the usage where there are two sorts of ways a theory can be epistemically or methodologically good: one way is in relation to evidence, and the other is by having structural or theoretical features such as simplicity, conservativeness, and the like. These people may well contrast “evidential” value with the sort of epistemic value I have in mind. My usage clearly differs from theirs, but so far as I can tell this difference is merely terminological.

Epistemic conservatism is common, though it is more commonly asserted to be good method (or presupposed to be good method) than argued for. Quine’s principle of “minimal mutilation” is in part a principle of theoretical conservatism—if new evidence requires me to shift the state of my beliefs, why ought I to make the minimum changes needed to restore coherence? Something like theoretical conservatism is also implicit in most Bayesian models of scientific inference. There are many admissible prior probabilities that one can have consistent with being rational, and they lead to different posterior probabilities when one processes evidence in the way Bayesians take to be rational. But even though this means that there is a range of posterior probabilities that are not ruled out by rationality, we are not rationally permitted to switch between these permissible posterior probabilities—we must stick to the ones we have, unless more evidence comes in that requires us to update our probability distributions. In effect, this means that what prior probabilities someone in fact has (or had) constrains what credences it is rational for that person to have: what the person does believe, by itself, makes a difference to what they should believe in a conservative way—they are entitled to stay put, but not entitled to change, unless updating in response to evidence. But theoretical conservatism is not simply a peculiarity of certain schools of philosophy of method—many “unaffiliated” (or less clearly affiliated) philosophers of science would agree with it.

At least one common methodology in philosophy in general seems to share in epistemic conservatism: the method of reflective equilibrium. While this method was also first articulated in the philosophy of science
(see Goodman 1955), it was of course popularized under the name ‘reflective equilibrium’, particularly in ethics and politics, by John Rawls (Rawls 1971). Reflective equilibrium is a process of beginning with our particular judgments about a topic, together with the generalizations we are initially inclined to accept, and rules of adjudicating disputes, and by repeated application of each to the others coming to a final theory in which inconsistencies and tensions between our starting commitments are all smoothed out. Goodman, and following him Rawls, thought that justification would thereby be conferred on our final resting place, though a lively debate can be had about why, and whether a commitment to reflective equilibrium as the model of methodology brings with it a pressure toward some sort of constructivism or other antirealism. (This debate is, I take it, a special case of the debate about whether coherentism in epistemology is in tension with realism in metaphysics.)

For our purposes, however, notice how reflective equilibrium suggests some form of epistemic conservatism. If one’s final point can be justified just because it comes from a starting point through a suitable process of harmonization, and where one ends up depends heavily on where one began, then plausibly one can become justified in believing \( p \) at a later time largely owing to the fact that one believed \( p \) at an earlier time. For example, it is quite plausible that two theorists could be rational and diligent practitioners of reflective equilibrium, and one be justified in believing \( p \) but the other not be, where the difference is entirely the result of the first including \( p \) in his starting commitments, while the second did not: the two could have starting points that are otherwise the same, and have done the same thing, mutatis mutandis, with their starting points. (Note well that this is a matter of epistemic justification, not merely some difference in appropriate pragmatic attitudes toward \( p \).)

It is hard for me to be more specific about the link between reflective equilibrium and epistemic conservatism, partly because the details of what reflective equilibrium is and what justifies its application are not matters of widespread agreement. (Of course we could just select a specific account of reflective equilibrium and examine how much epistemic conservatism it implied: but that would tell us more about the particular choice of formulation than the general link.) One thing in particular that might be relevant is what the “method” of reflective equilibrium tells us about justification during the process: we are told we have justification at the end of the process, but is there anything in general to be said about the justification of our beliefs at any particular stage in the middle of the process? This question deserves more attention than it often gets, especially since
in practice we are always in the middle of the process: nobody, not even Rawls, has completed the process of weighing up all our particular ethical judgments with all of our general principles and all of our methodological opinions about ethics. Nor did Goodman ever weigh up all of our particular judgments about inferences, our general inferential commitments, etc., to give a final answer about what “inductive” inferences we should engage in. Some tempting answers here, such as the claim we are justified in believing whatever has survived so far in the reflective process, at least absent evidence of serious error or incoherence, will give rise to fairly general principles saying that if we in fact believe \( p \), we have some reason to continue to believe \( p \) (perhaps with a few provisos to rule out justification in cases of particular epistemic depravity, or when a reasoner is sufficiently far from following the method of reflective equilibrium).

Epistemic conservatism deserves a defense in depth, especially given that many are suspicious of it, and some philosophers have argued in detail against it (for a recent survey and attack on epistemic conservatism, see Vahid 2004). That would take at least a paper in itself, so for now I will be content to associate the view with its partners in crime: those already tempted by Quinean webs of belief, Bayesian updating mechanisms, or the practice of reflective equilibrium should (nearly) all be tempted by epistemic conservatism in some form. A full discussion should also clarify exactly what sort of epistemic conservatism, if any, ought to be endorsed: in what circumstances does possessing a belief by itself oblige or permit an agent to continue to hold it? If it is a consideration to be weighed against others (as is surely plausible), how is this weighing to be done? These are all good questions, but they are all questions I want to sidestep here.

If epistemic conservatism is a virtue, then coming to realize what one believes is one way to come to realize what reasons one has for believing things as well: if the platitudes analysis reveals to me that I believe that properties are multiply located, then that shows me that I have some reason to think that they are.\(^{11}\) To the extent that epistemic method is not individualistic, similar things will be true in the case of shared beliefs, “common belief,” and the belief of experts: discovering that we believe something makes a difference to what we have reasons to believe. (Whether what we need to discover is what we each happen to believe in common, or something stronger like common belief, is a question about collective reasons for belief I need not tackle here.) We must be a little careful here—if having the belief is enough to ensure there is a reason, then the platitudes analysis does not generate that reason—it is already there; all the platitude
analysis would do in this case is *make us aware of* a preexisting reason for holding the beliefs in question. So the platitudes analysis does not give us these kinds of reasons (except insofar as putting commonplaces together may lead us to clean up our beliefs), but it only gives us explicit access to the reasons (or their grounds). However, if the epistemic conservatives are right, and having a belief is itself something that gives us a reason to hold that belief, then beliefs about whether or not one believes \( p \) themselves may give one an indirect reason for believing \( p \). This is because taking yourself to believe \( p \) is good, though perhaps not indefeasible, evidence that you do in fact believe \( p \). So the belief that I believe \( p \) is evidence that I believe \( p \), so it is evidence that I have evidence that \( p \). I am inclined to think that evidence of evidence is itself evidence (that is, evidence of evidence that \( p \) is itself evidence that \( p \)). This may not be obvious (in methodology, what is?) but some examples might help make it plausible:

1. I have a hazy memory that I proved Conjecture X to my satisfaction last week, though I cannot now recall the details.
2. A competent and reliable informant tells me that he has new evidence that Jones committed the murder, though he cannot tell me what that evidence is without breaking a confidence.
3. I have a visual impression of the shadow of a stick reaching the two-meter mark on a measuring tape. I happen to know the sun is at a 45 degree angle to where I am standing, so I take it that the stick is two meters long.

In each of these cases, I take it I have evidence of evidence of something—a memory of a deduction, testimony of some-evidence-or-other, and perceptual evidence of a measurement which itself is (in context) evidence for the height of the stick. In each of these cases I take it is quite plausible that the evidence of evidence for \( p \) is itself evidence: of Conjecture X, that Jones is a murderer, and that the stick is two meters long. If evidence of evidence were not itself evidence, then we would have to be very careful about many of our several-step chains of discovery.\(^{12}\)

If coming to believe that I believe that \( p \) is then itself evidence for \( p \), then the platitudes analysis can generate reasons to believe propositions about the world, rather than merely reveal that we have reasons we may not have noticed. I think this is one of the significant things going on when we engage in thought experiments and intuition pumps whose conclusions are not just about our concepts, but about how the world is.

Some might think that if epistemic conservatism has the result that getting new evidence for the world is this easy, that is a reductio of
epistemic conservatism. I do not intend to defend epistemic conservatism in depth here, but it is worth pointing out that this need not be very strong evidence at all: for all I have said here, it need not ever be enough to justify confident belief on its own. Even if it is not, it may serve as a valuable methodological starting place—if we are permitted to continue in our beliefs until we work out a better alternative that may be preferable to having to engage in widespread suspension of judgment. A commitment to epistemic conservatism, even with the additional commitment that evidence about our own beliefs is itself evidence of the truth about the subject matter of those beliefs, need not bring with it the suggestion that this is a particularly fruitful way to proceed in practice—I suspect it will be of most use at early stages of inquiry, before we engage sufficiently with a question to be able to do better than rest content with our starting place. It may also be helpful to point out the analogy with testimony—the discovery that others believe that $p$, for example, by their testifying that $p$, is often some evidence that $p$, even if that evidence is eminently defeasible and may count in favor of $p$ only because of contingent things we assume about the reliability of other people. And I suspect this reliance on others is fairly fundamental—it often, of necessity, provides the starting point of our opinions about the world rather than its being the case that we only take others’ opinions into account after carefully examining their bona fides and independently established reliability. Is it so terrible to extend to ourselves the same courtesy that we extend to the woman or man on the street?

Finding out what we think may serve one more related goal, but one that is probably less valuable. Contemporary philosophy is sometimes conducted in an adversarial manner, where we see the job of a view’s proponent to make the most persuasive case possible. Other times we have some tendency to treat it as a rhetorical exercise, where the job is to convince our readers or interlocutors of a view (or more feasibly, to get them to find the view more plausible than they did before). In this environment, we may end up just trying to convince people of our views, and lose sight to some extent of the question of what we have good reason to believe. So sometimes, I imagine, discovering what someone believes, or what we collectively believe, is done only to make someone consciously convinced of something: we may on occasion use an intuition pump or extract a consequence from folk theory, not just because we are aiming at getting a better theory, but simply to convince an interlocutor (or reassure ourselves).

I have suggested that when this is going on, often we are contributing to our theory’s value—either evidentially, or perhaps by improving some
pragmatic virtue. But this use is primarily ad hominem: we are merely trying to find beliefs that the party-to-be-convinced holds, or can be convinced to hold, that can be used as premises—and trying to do this merely because it is a way to get the person in question to agree with us. This is dialectically useful, and no doubt most of us stoop to this at some stage or another—but simply persuading someone that one’s view is correct is not necessarily very valuable unless one is bringing out reasons to believe the view being supported. (Not valuable if we are trying to find the truth, that is—there are other motives for trying to get people to agree with you, of course.)

I expect the opponents of epistemic conservatism will be more inclined to think that this is all that is going on when we are articulating views we may be holding implicitly, or make appeals to others by bringing out features of “common belief” or features of the views they implicitly hold. I think we should be on our guard, when presented with arguments from premises we accept, to make sure we have good enough reasons to believe the premises before we become too convinced of the conclusions. I can afford to be a little indulgent of my beliefs, given that I take epistemic conservatism to be a theoretical virtue, and an evidential theoretical virtue, at that: but those who do not had best be especially careful. Nevertheless, given that in fact we can secure agreement with another by convincing that person that they believed the proposition in dispute all along, no doubt the role that platitudes analysis can play in doing this is relevant to the philosophical enterprise as it is in fact practiced.

5 How the Platitudes Can Help: The Third Way

Discovering that one already has an opinion may in general be of some help in investigation—either because it is evidentiary, or for some more pragmatic reason. But it is particularly useful when one has some reason to suppose that people are good at coming to know the sort of thing the opinion is about. This applies to cases of conceptual analysis, presumably—one reason we think that analysis of concepts or expressions yields knowledge is that we have some useful access to our own concepts, or (perhaps more dubiously) to the meanings of expressions in public languages that we speak. If I find myself inclined to think that it is true in virtue of the meaning of the word ‘brother’ that anything that is a brother is a sibling, then part of why I take myself to be justified is that I think I know what ‘brother’ means in English. The claims that we know a lot about the concepts we use and the meanings of the words we use are both
controversial claims, and some theorists have denied one or both. But at least the traditional picture of conceptual analysis or linguistic analysis seemed to presuppose that we did reliably have such knowledge (though it was of course defeasible and perhaps limited in other ways).

We may have reason to suppose we are reliable about questions other than questions of meaning, of course. We may also on occasion have reason to think our pretheoretic judgments are to be trusted—or at least those judgments made before explicit, articulated theorizing are to be trusted, since we may have good reason to think our supposedly pretheoretic judgments are embedded in an implicit theory. Some questions in metaphysics might be thought to be like this—even if we find it very hard to articulate views or to know how to sort the true from the false ones, we might have reason to suspect we already know something about the subject that we are having trouble articulating, rather than thinking that we do not know what is going on at all. What it is to be an agent in the world, for example, is at least partly a metaphysical puzzle, and metaphysicians argue over whether we need to expand our metaphysical commitments to provide a satisfactory theory of agency. (Are an agent’s decisions after all some complex of physical and biological events, or something constructed from them in an innocent way? Or is there something fundamental and metaphysically special going on when an agent makes a choice that is absent from other phenomena?) Debates about freedom of the will, the metaphysics of mind and intentionality, the metaphysics of morals and normativity, and the like often touch upon this difficult-to-answer question. We might expect that we implicitly know quite a bit about agency—after all, we are agents, we are good at predicting and understanding the behavior of other agents (each other), and there is good reason to suppose that we even underwent evolutionary pressure to reliably track some things about agency—those without the self-awareness to monitor their decisions, or the social awareness to understand and anticipate others, will often have done less well. We might therefore take more seriously the deliverances of “common sense” about agency than we would or should take the deliverances of “common sense” about what happens to objects as they approach the speed of light or when they are one nanometer across.

Another example concerns abstract ontology. Before we do much explicit theorizing, we have a lot of beliefs about modal matters—we have beliefs about what is possible, we have habits of ascribing dispositions to things, and we are able to make conditional claims and take them to be justified or unjustified in particular cases, well in advance of being able to articulate general principles about them or to answer some of the questions we would
like general theories of modality to answer. Examples might include our conviction that if something necessarily occurs then it possibly occurs, or that discovering that something is in fact the case shows that it is possibly the case. On the other hand, our ordinary practices and ordinary ways of talking seem to have very little to tell us about what large cardinal claims in set theory are true (i.e., claims about how different very large infinite sizes of sets are related). A case can be made that we should take our platitudinous attitudes toward modality more seriously as evidence than we should take our initial opinions about large cardinal axioms in set theory.

So the third way the platitudes analysis might help us is when it reveals to us a theory that we have some independent reason to believe is likely to be reliable. We may in fact be in a position to take a theory to be justified or reliable even before we can properly say why it is reliable or why it gives us warrant—I presume we can trust the evidence of our eyes before we have a good story about the epistemology of perception, and we can often rely on the testimony of others before we know the details of how or why they come to be authoritative reporters. Something similar may be true when we come to answer questions in the metaphysics of agency or time or causation—I am inclined to think our implicit grasp of causation is better than our explicit ability to answer questions about causation, which is one of the reasons why I think intuition pumps and assembling platitudes about causation is a worthwhile exercise (though I am inclined to think that we will eventually be able to do better than present cleaned-up folk wisdom about what it is for one thing to cause another).

6 Are Any of These Three Things What People Are Actually Doing?

It is comparatively rare for people to explicitly take themselves to be employing a platitudes account when doing philosophical analysis. Many people are not even clear what they take themselves to be doing when they are writing about metaphysics—they may call something they put forward an ‘analysis’, or they may rely on an intuition pump, or they may offer a counterexample to a theory without being explicit about what their justification for taking the counterexample to be correct is. Whether in such cases people take themselves to be doing conceptual analysis, or applying the results of another investigation (e.g., a result from the natural sciences), or doing something else, is not always clear.

Though I would sometimes like people to be clearer, I do not even want to claim that people should always be. Not every paper is a methodology
paper, and if we insist that every time someone gives us an argument he also must give us the methodology behind that argument, we may end up in an unpleasant regress (e.g., if we expect people to justify their methodologies as well, and justify that justification, etc.). Or we may just end up with less satisfying work on the original problem—methodological exкурsus is sometimes distracting, and we are usually worse at saying why X is a reason for Y than telling whether X is in fact a reason for Y. So sometimes at least it is fine for a metaphysician (or philosopher, or other theorist) to just get on with doing what they are doing, and we can work out what their methodology was or should have been after the fact (if we bother to explicitly work it out at all).

So my guess is that some metaphysicians are “just getting on with it”: thought experiments, counterexamples, and plausible commonplaces are employed, whether they work through conceptual analysis, or a deployment of folk theory, or something else.\textsuperscript{13} I offer these three ways platitudes can help not as a story about what metaphysicians think they are doing, nor even, necessarily, as a story about what they are doing, though keeping them in mind while answering that question may be helpful. Instead, they are offered as justifications for practices that many of us are engaged in.

The alert reader will have noticed that my remarks about metaphysics and the Canberra Plan have wider application than to the activity of Canberra Planners and fellow travelers. Anyone interested in fundamental metaphysics and interested in methodology should be interested in the question of how armchair methods can help us. Someone can disagree with various parts of the Canberra plan project, or alternatively be relatively uninterested in where our everyday notions fit into the world as described in more fundamental vocabulary, and still see the merit in the three ways discussed that the armchair can help in fundamental metaphysics. Indeed, people who are not even particularly interested in metaphysics might find this discussion useful, since virtually no matter what inquiry one is engaged in, the question of how “relatively a priori” investigation can help may be interesting. Perhaps these sorts of contributions in some areas are overshadowed by other methods: but investigating the world can be hard, and it would be a shame to neglect any source of epistemic justification we have available.

Acknowledgments

Thanks to Chris Daly, Alan Hájek, Katherine Hawley, Ishani Maitra, and audiences at the 2003 Metaphysics in Science Workshop in Athens, the
2003 AAP conference in Adelaide, and at the Universities of Glasgow and Leeds, La Trobe University, and UC Davis, for comments and discussion. Special thanks to the staff and students in the Philosophy Program at ANU from 1994 to 1996, for Planning.

Notes

1. The expression is originally due to O’Leary-Hawthorne and Price 1996. A more recent appearance of the expression in the literature can be found in Lewis 2000a.

2. The expression is from Jackson 1998a. Earlier in “Armchair Metaphysics” Jackson referred to these as “placement problems” (Jackson 1994a), and this terminology has been picked up by others, e.g., Price 2004.

3. There are other variants along these lines—instead of listing all of the relevant sentences we take to be true, or all the relevant sentences in a theory, we may instead wish to list a set of sentences which constitute some privileged organization of the relevant sentences (e.g., a set that jointly entails all of what we believe).

4. I take an example of this to be the use of ‘platitudes’ in Smith 1994, 29–31. A related usage is that of Wright in, e.g., Wright 1998. Wright is not a Canberra Planner, but his usage is obviously related (and he tells me the relation is ontogenetic as well as phylogenetic). Examples could be easily multiplied.

5. This would not be the influence intended by these authors, if we took platitudes to capture the meaning of theoretical terms, of course.

6. One thing it might be talk of is only changing what we take to be the platitudes associated with a term. Although strictly speaking this would not be changing the platitudes but only changing our beliefs about them, English does permit this way of speaking (perhaps it is metaphorical). It would be similar to the way of speaking in which we can appropriately say that nineteenth-century physicists changed the electron several times, or a modern archaeological discovery changes the ancient Egyptians from an insular culture to one embedded in very long-range trading networks.

7. This example is merely illustrative—some might think that it will turn out to be the property of having a certain ancestor, or will involve the possibility of interbreeding, or something else. This is not intended as a serious conjecture in the philosophy of biology.

8. I don’t mean to suggest that the people I talk to about philosophy are a subset of the people who agree with me—if that were the case, I wouldn’t have philosophical conversations very often!

9. Most famously outlined in Quine 1951, though he does not use the expression ‘minimal mutilation’ there.
10. One broadly Bayesian view that is an exception is Bas van Fraassen’s voluntarism—see, e.g., van Fraassen 1989, chapter 7 for a discussion.

11. That properties are multiply located is not in fact something I think a platitudes analysis would reveal is part of a folk theory of properties.

12. Exercise for the reader: see if you can come up with cases where evidence of evidence that $p$ is not itself evidence for $p$.

13. Another thing that I think one can do from the armchair is apply theoretical virtues such as simplicity, unification, and perhaps some others: many of these can be subsumed under a category of considerations that make for a better explanation, in the “inference to the best explanation” sense. Some claim that these virtues are only pragmatic: I am inclined to think many are evidential, but in either case as long as theories can be improved by their application, applying these standards (e.g., by arguing some alternative is simpler or more unified than another) can be worthwhile—and applying these standards can often be usefully done without additional data collection.
III  Normativity
13 Naturalizing Normativity

Mark Colyvan

1 Introduction

Normativity is built into the very notion of rationality. For example, when we describe some set of beliefs or actions as rational, we mean that these beliefs or actions are sanctioned by the relevant (normative) theories (Bayesian belief theory and decision theory, respectively). We might add logic, the study of rational inferences, to the mix. Indeed, logic, belief theory, and decision theory are quite unlike other scientific theories in this regard. These three theories do not purport to simply describe or systematize the way real-world agents reason and act; these theories purport to describe and systematize how idealized rational agents reason and act. Or, to explicitly invoke the normativity here, they purport to prescribe how real-world agents ought to reason and act. But this raises an important question about where such theories derive their normative force.1

The problem is that normative operators such as ‘ought’ are rather odd modal operators—they are not truth-functional2 and, as David Hume pointed out (Hume 1975 [1747], 469), one cannot derive normative claims from matters of fact. Yet, ‘one should ensure that one’s beliefs accord with Kolmogorov’s axioms’ was not given to us by divine decree. In this paper, I propose one way to understand the origin and force of the normativity of theories of rationality. I do this from a broadly naturalistic point of view. Although from this point of view, normativity is returned to its rightful place in center stage, the advice delivered about rationality is not the usual advice. I argue that the price we pay for normativity is a kind of fallibilism, and this opens the door on alternatives to classical logic, Kolmogorov probability theory, and standard decision theory. Finally, I compare this approach with another naturalistic approach to normativity proposed by Robert Nola 2003 and Frank Jackson 1998a.
2 From the Nonmodal to the Modal

It might seem clear that one cannot get a normative claim from a descriptive claim. For instance, people use modus ponens in everyday reasoning, but from this it does not follow that people ought to use modus ponens. Indeed, it would seem that it is not just normative notions that are problematic; there is a broader class of modal notions that resist derivation from matters of fact. After all, modal notions like necessity seem to suffer similar problems: the modal proposition $\Box P$ does not follow from the nonmodal matter of fact $P$. But in the more general modal case, at least, it would be rash to conclude that modal notions cannot be derived from descriptive claims. It’s true that you can’t get $\Box P$ from $P$, but from a descriptive claim about how all the relevant possible worlds are, we can most definitely derive modal claims. So, for example, from the descriptive claim ‘for all worlds $w$, $P$ is true in $w$’, we can legitimately derive the modal claim $\Box P$.

Invoking possible worlds semantics for modal claims to move between the descriptive and the modal might be seen as something of a cheat and irrelevant to the main issue. After all, it might be argued that unless one is a realist about possible worlds, claims about possible worlds other than the actual are not really descriptive claims at all, but, rather, are just disguised modal claims. I won’t take a stand on this issue here; it is not my purpose to defend the more general claim that all modal notions can be derived from matters of fact. I will be content to demonstrate that various normative claims concerning rationality can be derived from matters of fact. Why then raise the issue of the modal semantics? Because it demonstrates that often we are quick to think of modal claims as being of a quite different kind from nonmodal claims and a case can be made, at least, that this is mistaken. Once one looks for the right kind of nonmodal facts, the apparent gap between modal facts and nonmodal facts disappears. So too, I will argue, is the case with claims about rationality.

3 From the Descriptive to the Prescriptive

Consider some typical prescriptive claim concerning rationality:

(1) It is rational to use modus ponens.

Now consider some descriptive claim such as

(2) Agent $A$ employs modus ponens.
Clearly (1) does not follow from (2), because, apart from anything else, A might be a poor reasoner, and quite unjustified in her use of modus ponens. But (2) is not the only relevant descriptive claim. Let’s look further afield (as we did in the last section when looking for descriptive counterparts for modal claims). Consider

(3)  Most agents employ modus ponens.

Does this help in deriving (1)? No, for much the same reason: most agents could be bad reasoners. But what if we restrict attention to good reasoners? Consider

(4)  Good reasoners employ modus ponens.

There does not seem to be such a gap between (1) and (4), but now another problem emerges, namely, circularity. The phrase ‘good reasoners’ in (4) is very suggestive of ‘rational agent’, so we are a small step away from providing a circular definition of rationality, in terms of what rational agents do. It would seem that we’ve closed the normative–descriptive gap, but at the price of rendering the account circular, and thus useless.

But if we can give an independent account of what a good reasoner is, we can avoid the charge of circularity. I claim that we can do this, but to see how, it will be useful to look at a related example. Consider the concept of ‘a good chess move’ in a given setup. Obviously a good chess move must be a legal move, but it must be more than this. We might define such a move in terms of what good chess players would do in such circumstances. Our definition of ‘a good chess move’ also seems to be circular—at least until we can spell out what ‘a good chess player’ is. But the latter is easy. It does not have to be spelled out in terms of someone who makes good chess moves; we can spell it out in terms of someone who wins games of chess. Or better still, we can spell out what a good chess player is in terms of their moves according well with the relevant bits of chess theory. With this example in mind, let’s return to the question of rationality.

One way to answer our question of what a good reasoner is is to provide a teleological account according to which a good reasoner is one who is better adapted to achieving certain goals. There are a number of candidates for the appropriate goals. The most popular of these is survival. While such evolutionary accounts of rationality undoubtedly have something to recommend them (see, e.g., Nozick 1993; Papineau 2003), typically such accounts are too undiscriminating. After all, not all poor reasoners have been selected against. A better answer to the question of what a good reasoner is is that it’s someone whose reasoning is well supported by our best
theories of reasoning. Again, it might seem that we are dangerously close to circularity, but, be that as it may, circularity can be avoided. Our best theories of rationality—formal logic, belief theory, and decision theory—are not defined purely in terms of what good reasoners do. These theories earn their place as our best theories of rationality in much the same way that general relativity earns its place as our best scientific theory of space and time.

There is another feature of theories of rationality that deserves mention. Theories of rationality must employ a great deal of what they set out to explain and systematize. Take, for example, a debate in logic—the debate over the validity of excluded middle, say. In order to settle the issue of whether we ought to employ a logic in which excluded middle holds (such as classical logic) or one in which it fails (such as Kleene strong $K_3$) we need to reason and, in particular, we need to make, and assess the validity of, various logical inferences. But which logic do we use for this task? Philosophical logic, like other subjects that use the very theories under study, is difficult, but not impossible. Indeed, it would seem that the three theories of rationality that I have in mind for this discussion—formal logic, belief theory, and decision theory—all have this reflexive character. This makes their study difficult, but it does not undermine their value or preclude progress on the study of rationality.

But putting aside the worries about the reflexive nature of the rational study of rationality, another more pressing problem emerges. So far, I’ve argued that we can get from descriptive statements to normative ones by invoking a certain respect for scientific theories. But on what grounds can this be justified? Indeed, this move, once spelled out, just smuggles normativity in through the back door. For surely the principle I’m appealing to is some version of the doctrine of naturalism (see Quine 1969) that counsels us to believe our best scientific theories. Or, to make the normative assumption manifest, it’s rational to believe or accept our best scientific theories. I agree that naturalism (of the kind I subscribe to; see Colyvan 2001, chapter 2) is normative, but I deny that this undermines the account of the normativity of rationality I’ve given so far. I discuss this in the next section.

4 The Role of Naturalism

As we saw in the previous section, my justification of the normativity of various theories of rationality hangs crucially on the doctrine of naturalism. We’ve also seen that this doctrine is normative—it tells us that it is
rational to believe (only) our best scientific theories. So the cogency of my argument hangs on a defense of naturalism. Unfortunately, a defense of such a fundamental doctrine as naturalism is hard to come by. One finds different positions on this issue. Some suggest that naturalism is just an attitude one takes toward investigating the world. It’s just a basic commitment to attempt to describe and explain the world without resorting to supernatural or otherwise spooky entities and forces. But I think we can do better than simply taking a stand. I suggest that the naturalistic approach has many benefits and we should judge the doctrine by its fruits, not by arguing for it from more fundamental principles.9

So what are the fruits of naturalism? First, the scientific enterprise has a remarkably successful history, and naturalism is little more than a statement of our continued support for that enterprise. After all, rejecting naturalism amounts to claiming that sometimes we ought not accept our best scientific theories. Let’s get clear about what this amounts to in the current context. With Quine, I’m understanding science very broadly here, to include all theoretically and empirically well-supported areas of study (including philosophy). In short, our best scientific theories are simply our best theories. To reject naturalism is to deny that we ought to accept our best theory of some domain. But what are the other options? Accept the second-best theory? Accept no theory at all?10 Once put this way, naturalism, if not self evident, is at least a rather compelling doctrine. It’s not trivial though. It does rule against certain mystical and religious world-views, for instance—at least when there are better (scientific) theories of the same phenomena. Other benefits of naturalism are that it enables a rather plausible reply to the skeptic (see Quine 1974, 3) and it provides a satisfying account of the relationship between philosophy and science.

Once again, I stress that although the above justifications of naturalism are circular, they are not viciously so. All chains of justification ultimately result in either infinite regress or circularity. All other things being equal, circularity seems the more palatable option, so long as the circles are not too tight (i.e., vicious).11 If all I’ve said so far is correct, we have a sketch of an account of the normativity of rationality theories. This account is based on the acceptance of the naturalistic point of view. There are, however, no free lunches. Let’s consider the consequences of this particular defense of normativity.

The account of normativity I’ve been defending depends on naturalism’s respect for the (current) best scientific theories of rationality. It is thus a defeasible account, and typically canons of rationality will change depending on the state of science at the time. What counts as a valid inference
at one time may not count as a valid inference at some other time. Those who would have classical logic, standard Bayesian belief theory, and standard decision theory as the normative theories of rationality may be uncomfortable with this result.12 For, on the view I’m advancing here, these theories are only defeasibly the normative theories of rational inference, rational belief, and rational decision making, respectively.13 But seen in the right light, this defeasibility is both natural and desirable. Our current best scientific theories of rationality should not be any different from other scientific theories. All these theories are known a posteriori, they are defeasible, and their acceptance or rejection is sensitive to new evidence and new developments.

5 The Ramsey-Lewis Approach to Normativity

Let me finish up by saying a little about the points of contact with, and the differences between, the account I’ve sketched here and the Ramsey-Lewis approach (see Ramsey 1990, chapter 6; Lewis 1970) to defining theoretical terms (otherwise known as “the Canberra Plan”). The Ramsey-Lewis approach provides a rigorous means of providing implicit definitions of theoretical terms by appealing to their role in the theories in which they occur. Suppose, for ease of exposition, that there is only one theoretical term, $\tau$, we are hoping to define. We separate the theory in question, $\Gamma$, into theoretical terms (which in this case is just $\tau$) and nontheoretical terms, $d_i$. The resulting theory can be written as

$$\Gamma(\tau, d_1, d_2, d_3, \ldots, d_n).$$

We then existentially generalize to get the Ramsey sentence

$$(\exists x)\Gamma(x, d_1, d_2, d_3, \ldots, d_n).$$

Finally, we replace the existential quantifier in the Ramsey sentence with the definite description operator ($\langle x \rangle$) to obtain our implicit definition of $\tau$.

$$\tau = \text{Defn} (\langle x \rangle)\Gamma(x, d_1, d_2, d_3, \ldots, d_n).$$

This can be generalized to cases of $m$ theoretical terms, $\tau_1, \tau_2, \ldots, \tau_j, \ldots, \tau_m$, with the corresponding Ramsey sentence:

$$(\exists x_1, x_2, \ldots, x_j, \ldots, x_m)\Gamma(x_1, x_2, \ldots, x_j, \ldots, x_m, d_1, d_2, d_3, \ldots, d_n).$$

We can then provide implicit definitions for each $\tau$ in terms of the non-theoretical terms and other theoretical terms:14
In the case of theories of rationality, the relevant theories will be Bayesian belief theory, deductive logic, and rational choice theory. Jackson (1998a) and Nola (2003) have shown how to extend the Ramsey-Lewis approach to normative theories such as ethics and rationality (respectively). As Nola (2003, 162) points out, all we need to do is separate the normative terms (such as ‘rational’) from the nonnormative or descriptive terms of the theory. The resulting implicit definition of the normative terms in question appeals only to their functional role in the theory and this is articulated only in nonnormative vocabulary. There is a sense in which this bridges the gap between the normative and the descriptive in a naturalistically respectable way. More on this shortly.

It might be thought that the Ramsey-Lewis account is simply a formalization of the approach I’ve argued for in the earlier sections of this essay. There are certainly some similarities between the two approaches. Both take naturalistic approaches to dealing with normativity and both focus attention on theories of rationality. And they both agree that there is something unsettling about normative notions. The Ramsey-Lewis approach’s response is to eliminate the normative in favor of the nonnormative. The approach I’m suggesting reduces all normativity to just one—the normativity implicit in naturalism. But both accounts are naturalistic and they both take it that normativity deserves to come in for special treatment. That is, they both set out to address the Humean worry about deriving normative claims from descriptive claims.

Despite these initial similarities, there are also some important differences between the two approaches. Take rational choice theory, for instance. Here the normativity—what an agent ought to do—is spelled out in nonnormative terms—the maximizing of expected utility. Both approaches seem to be on the same page so far, but notice a couple of important differences. In the Ramsey-Lewis approach (at least the version advanced by Nola 2003) the normative–descriptive gap is bridged by providing a definition of normative terms in terms of purely descriptive terms of the relevant theory. The approach I’m advocating doesn’t claim to provide anything so strong as a definition of the normative terms in question. After all, nowhere is it claimed that the meaning of ‘what you ought to do’ is ‘maximize expected utility’. So there’s one significant difference between the two approaches: the Ramsey-Lewis approach invokes implicit
definitions to bridge the normative–descriptive gap; the approach I’m defending does not offer definitions at all.

There is a second, more significant difference between the two approaches. The Ramsey-Lewis approach defines the normative terms via their functional role in the relevant scientific or folk theories. My approach, on the other hand, uses the scientific theory to back up the normative claims in question: you ought to do what your best theory says to do. Why? Because you ought to believe your best scientific theories. And the threatening regress is blocked by invoking the doctrine of naturalism. In effect, on this approach, all normativity arises from the normativity built into the doctrine of naturalism. Of course the Ramsey-Lewis approach is also naturalistic. But it is naturalistic in a quite different way. The naturalism doesn’t provide the source of normativity; the Ramsey-Lewis approach is naturalistic in the sense that normative terms are defined via naturalistically respectable descriptive terms. In the Ramsey-Lewis approach the aim is to arrive at a naturally respectable theory—one without suspect terms—but naturalism does no real work in the account. Naturalism provides only the motivation. On my account naturalism does some serious work, for it is naturalism that is the ultimate source of the normativity in question.

The final difference between the two approaches concerns the scope of their applicability. It is not clear that the approach I’ve argued for can be extended to areas where there is no well-defined, consistent, and unique formal theory. My approach can only by applied to well-systematized (indeed, scientific) theories of rationality. The account works well for the theories I’ve been explicitly discussing—formal logic, belief theory, and decision theory—but it’s not at all clear that the account presented here can be used for less rigorous theories of rationality (such as commonsense judgments and informal logic). For similar reasons, extending this account of normativity to ethics may also prove problematic; it’s not clear that there is a suitably rigorous and systematic theory of ethics on which the account can rely. The strict version of the Ramsey-Lewis account may too suffer such limitations, but it is generally taken to be a core part of the Canberra Plan that the Ramsey-Lewis approach to defining theoretical terms can be extended to folk theories. But in any case, I see no prospect for the approach I’ve argued for to be extended in this way. This would thus seem to mark a significant difference between the two approaches.

This brings us to the question of which is the better approach. Indeed, in light of the last point, the wider scope of the Ramsey-Lewis approach would suggest that the latter has the edge. But I think that it is a mistake to compare the two in this way. There is a sense in which these are two
completely different games. The Ramsey-Lewis approach, motivated by a
naturalistic suspicion of normativity, aims to dispense with normative
terms in favor of nonnormative terms. As I’ve already mentioned, though,
in this approach, naturalism plays no role beyond motivation. The Ramsey-
Lewis approach seeks to satisfy the exacting standards of a certain concep-
tion of naturalistic philosophy. The approach that I’m advocating employs
naturalism to defend the normativity so central to theories of rationality.
The Ramsey-Lewis approach is apologetic for normativity and, in the name
of naturalism, seeks to do away with normativity. My approach embraces
normativity and employs naturalism to allow normativity to take its right-
ful place in theories of rationality. Once put this way, the question of
which approach you should prefer boils down to a question of what you’re
trying to do and how you conceive of the naturalistic project of philoso-
phy. Indeed, it may well turn out that these two approaches to normativity
are not competitors at all, just different parts of, or different takes on, a
broader project of naturalizing normativity.

Acknowledgments

An earlier version of this essay was presented at the joint Fourth Interna-
tional Conference on Cognitive Science and the Seventh Australasian
Society for Cognitive Science Conference at the University of New South
Wales in July 2003 and in a joint Philosophy and Science Studies Seminar
at the University of South Carolina. I’d like to thank the audiences at those
presentations for their contributions. I am also grateful to Damian Cox,
Katie Steele, and Juliana Weingaertner for helpful discussions, and to Alan
Hájek and Carol Booth for comments on an earlier draft.

Notes

1. There is, of course, a place for purely descriptive theories of human reasoning,
but these are not the focus of this essay.

2. For example, the truth of ‘you ought not affirm the consequent’ cannot be deter-
mined purely from the truth of ‘you don’t affirm the consequent’.

3. Although the modal $\Diamond P$ does follow from the descriptive $P$, at least under some
interpretations of ‘$\Diamond$’.

4. See Hintikka 1962 for a classic formal account of knowledge and belief in terms
of modal operators.

5. Indeed, anyone who has ever taught introductory formal or informal logic will
know that a large number of incoming students take various fallacies (e.g., affirming
the consequent) as valid argument forms. Wason and Johnson-Laird 1972 explores these issues.

6. I won’t go into the details here. Elsewhere, in Bueno and Colyvan 2004, I’ve argued that even apparently fundamental principles such as the law of noncontradiction can be sensibly debated without begging questions.

7. After all, psychology manages to use the human mind to study the human mind and physics employs physics to study physics. We should not become what Adam Elga calls ‘reflexophobic’. That is, we should not be so struck with the various problems and paradoxes of self-reference that we shy away from self-reference wherever it occurs. Much self-reference is benign.

8. I hedge by including both ‘accept’ and ‘believe’ here so as not to beg any questions against constructive empiricists such as van Fraassen 1980.

9. This is the same way modern set theorists argue for the acceptance or rejection of new set-theoretic axioms. See Russell 1973 and Gödel 1983 (1947) for early advocates of this modern view of the foundations of mathematics.

10. Naturalism, properly construed, doesn’t rule against an occasional bout of agnosticism. For instance, naturalism may license agnosticism when there is no best theory or when the best theory is clearly inadequate.

11. It has been claimed by Katz 2000 that the kind of naturalism under consideration here is worse than circular—it leads to paradox. Elsewhere (Colyvan 2006), I’ve argued that Katz is wrong and that no paradox can be generated—at least not without making some rather implausible assumptions about the logic of belief revision.

12. For example, a great deal of work on rationality in psychology takes these classical theories of rationality to be the final word on normative rational theory. Kahneman and Tversky (see many of the essays in Kahneman, Slovic, and Tversky 1982) take standard Kolmogorov probability theory to be (uncontroversially) the theory of belief and belief updating. And Wason and Johnson-Laird 1972 take classical logic as the theory of deductive inference. See Hájek 2003a,b for some of the shortcomings of classical (conditional) probability theory and Shafer 1976 and Walley 1991 for examples of nonclassical belief theories.

13. In fact, I think classical logic has already passed its use-by date. And insofar as standard Bayesian belief theory and standard decision theory presuppose classical logic—it’s the tautologies of classical logic that are defined to have maximal probability in Kolmogorov probability theory—those theories too are questionable. See Colyvan 2004 for more on this issue.

14. The complications are a result of defining one theoretical term by way of other theoretical terms, where the latter are to be defined in the same manner. These complications need not concern us here. See Lewis 1970 for details.
15. See Colyvan, Cox, Steele (forthcoming) for some steps toward formalizing ethical theories.

16. Although I won’t enter that debate here, there is some reason to be suspicious of such extensions. For the Ramsey-Lewis approach to work, we need to have something that resembles a theory, and folk “theories” are often inconsistent, or at least nonunique and usually imprecise.
14 Moral Functionalism, Ethical Quasi-Relativism, and the Canberra Plan

Denis Robinson

1 Agenda

In this essay I pursue a double agenda. I believe the items on that agenda complement one another. One item is to illuminate some issues relating to the Canberra Plan. My main focus will not be on generic issues, but on matters specific to the distinctively evaluative and normative domain of ethics. The other item is to promote a view—more properly, a sketch of a view—I favor, which I dub ethical quasi-relativism.1 I shall discuss, as primary representative of a “Canberra Plan” treatment of ethics, Frank Jackson’s version of “moral functionalism,” especially as it is set out in his From Metaphysics to Ethics (Jackson 1998a).2 My discussion will follow a somewhat zig-zag course as I compare my view with Jackson’s, and consider various issues that are relevant to their similarities and differences. My thought is that a view like mine emerges fairly naturally if we take a paradigmatically “Canberra Plan” view such as Jackson’s, acknowledge certain problems for it, and revise our view accordingly. If that thought is even near the mark, each view stands to be illuminated by the comparison.3

The Canberra Plan typically operates within an implicit framework that views ordinary speakers of some problematic discourse as holding a straightforwardly descriptive theory of the world and how it works, implicit in that discourse, and seeks to determine whether and under what constraints the world does, or even could, conform to such a theory. Clearly a Canberra Plan treatment of ethics (much as with any descriptivist treatment of ethics) will confront the question of how to accommodate the evaluative and normative dimensions of ethical discourse. One main focus of debate must be the “Open Question” argument, and the difficulties of inferring ‘ought’ statements from ‘is’ statements. Another must be the “internalism”
constraint, which insists that there is a necessary connection between moral judgment and motivation. Although Jackson has interesting and illuminating things to say about these, I shall not say much about them. A third focus—which will be the principal focus of my discussion—is on the role of assertions and denials of moral claims, in giving expression to certain kinds of disagreement.

When we try to characterize moral disagreements, a special subset of them (at least) proves difficult and puzzling. I shall defend a view that attempts to make room explicitly for disagreements of this distinctive and puzzling kind. I will claim that to do so, we must take a view of how moral concepts are to be described and individuated that is more sophisticated than the view a standard Canberra Plan approach, such as Jackson’s, makes available.

2 Relativism, Disagreement, and Failures to Disagree

I'll start by considering some issues about agreements and disagreements. What I'll call “simple” relativism about \( X \) says that (i) truths about \( X \) are relative to agents or standpoints, and (ii) that apparent disagreements about \( X \), where due to different standpoints, are therefore really nondisagreements. Quasi-relativism about \( X \) differs (in a way) about (ii): hence the prefix ‘quasi’. Thus ethical quasi-relativism claims differing concepts of right, wrong, good, evil, permissible, impermissible, etc., may be legitimate—unfaithful neither (a) to facts nor (b) to established usage, nor (c) to other legitimate constraints, at least by non-question-begging standards (the no-fault condition). But it claims that nevertheless, from such divergences, a kind of bona fide disagreement springs (the bona fide disagreement condition).

But there is a puzzle in how the no-fault and bona fide disagreement conditions can be met simultaneously. Indeed on too narrow a view of what counts as disagreement, I don’t think they can be. To progress, we need to examine not only disagreements, but also what I’ll call ‘failures to disagree’. To this end I’ll start with some illustrative dialogues between characters named ‘Argle’ and ‘Bargle’. (Think of these as surnames shared by members of two argumentative clans.)

First, then, some failures to disagree, or, as I’ll call them, ‘mere argie-bargie’. Frank Ramsey left us a benchmark case (see Ramsey 1990, 247) that sets some kind of relatively low point on a scale of perplexingness:
Ramsey’s case

Argle: I went to Grantchester this afternoon.
Bargle: No I didn’t.

This case hinges entirely and obviously on the phenomenon of indexicality, which most of us understand better than Bargle apparently does.

“Delicious” dialogue

Argle: Yum, this Vegemite is delicious.
Bargle: No it’s not.
Argle: It is to me!
Bargle: Well not to me.

This case is broadly similar to the previous, though the indexical element is not initially overt.

Sofa and divan dialogue

Argle: You said they’d given us a room with a sofa, but there’s only a divan!
Bargle: But a divan is a sofa!
Argle: Nonsense! A sofa has arms and a back.
Bargle: On the contrary, a divan is precisely, a sofa which lacks a back, and possibly also lacks arms. This one has arms though. Don’t be fussy!

Argle and Bargle associate different application-conditions with the word ‘sofa’. Understood Argle’s way, ‘the room has a sofa’ is false; understood Bargle’s way, it’s true. They are each right in their own way and there is no real contradiction in their claims.

First football dialogue

Argle: In football, it is permissible to pass the ball by punching it.
Bargle: It is not!

God (thinks): Poor mortals! Argle is thinking of Australian Football, Bargle of soccer, so they are invoking football-concepts with different application-criteria, hence expressing no genuine disagreement.

Bank dialogue

Argle: After Harry arrived in River City he went down to the bank.
Bargle: Well I was there, picnicking down by the river, and I didn’t see him.
Argle: Of course not: I meant the savings bank, not the river bank.
I see all these cases as straightforward failures to disagree. What moral can we draw?

When protagonists make claims having truth conditions that are mutually logically inconsistent, let’s say they have a ‘basic logical disagreement’. It’s easy to think, real disagreement requires basic logical disagreement: and this seems to fit the above cases. Token utterances superficially appear in each case to be contraries—to have jointly unfulfillable truth conditions—but in the mouths of their utterers, they do not. Either difference of context (including speaker) makes for differing truth-conditions, or the same or superficially similar words are used with different application conditions (we could say, to express different concepts). So no basic logical disagreement; so no real disagreement.

But is it always true that where there is no basic logical disagreement, there is no “real” disagreement? My aim is to argue otherwise. At least, I’ll argue that there can be cases where there is “a kind of disagreement” despite a lack of basic logical disagreement. (I mean here to suggest the colloquial answer “kind of” to the question ‘is there a real disagreement?’.

“Merely semantic” debates about proper use of the phrase ‘real disagreement’ are not what concern me here.)

Let’s widen our range of examples. In particular, let’s ask whether the following is a “failure to disagree”.

Euthanasia dialogue

Argle: Assisted voluntary euthanasia violates my entire concept of what’s right, which holds human life sacrosanct. It is always wrong.

Bargle: Au contraire, it’s often right. According to my concept, what’s right is what maximizes well-being and minimizes suffering, so long as certain core moral values are respected. Autonomy is one of them. Coupled with the imperative to minimize suffering, it trumps the so-called sanctity of human life.

Viewed as we viewed the previous examples, this too should be a case of mere argie-bargie. But intuitively it is not!

Of course a great many doctrines on the philosophical market offer to explain this fact. But let’s for now attempt to simulate a kind of pretheoretical innocence and see where it leads us if we try to take these statements—including the references to “concepts”—at face value (trying as we do to begin also with a presumption in favor of some kind of ethical cognitivism). I will suggest that sometimes ethical disputes really do have a character of the kind reflected in this dialogue, so viewed. There is a discrepancy
between moral concepts deployed by the disputants, so that there is no basic logical disagreement: yet they still count intuitively as a “kind of” disagreement.

I hold in addition that in some such cases, it is idle to ask which party has, in absolute terms, the correct view, as if this question can be answered from some neutral standpoint, since there is no suitably “neutral” standpoint. Any attempt must import a nonneutral standpoint and hence be essentially question-begging. For want of a better term, I’ll stipulatively call these disputes ‘irresolvable’; henceforth I’ll use the term quasi-disagreements for irresolvable disputes that involve no basic logical disagreement but which stubbornly resist being viewed as nondisagreements or mere argie-bargie.

3 Moral Functionalism—Whose Theory? Which “Folk”?

Let’s now consider a few more details of, and issues for, Jackson’s moral functionalism. Moral functionalism is named for its analogy with analytic functionalism in the philosophy of mind. In a nutshell, the latter suggests that users of commonsense or “folk” psychological vocabulary tacitly subscribe to a theory, “folk psychology,” seen as implicitly defining terms of that vocabulary in nonmentalistic, commonsense terms. Analytic functionalists hold that were such a theory to be made explicit, purged of inconsistencies, and Ramsified, it would be appropriate to regard the (folk) mental properties as the (joint, near enough) satisfiers of the resulting Ramsey sentence. Moral functionalism is similar except that the relevant theory is folk morality, seen as implicitly defining moral terms in nonmoral descriptive commonsense terms. Moral properties are thus seen as the best (joint, near-enough) satisfiers of the folk-moral roles specified by the relevant Ramsey sentence.

In support of this approach, Jackson begins by arguing for cognitivism and descriptivism. Indeed he presents his moral functionalism as the working out of what he calls ‘analytic descriptivism’. He argues for the latter from what he claims to be the a priori, necessary, global supervenience of the ethical on the descriptive.

Let’s look more closely at the question: how, on views like Jackson’s, do we build (or contemplate building) a folk theory, and out of what? Appeal to intuitions about actual and possible cases (the “method of cases”) is not in principle, nor historically, the only option. One option suggests that we take as raw materials for a folk theory a suitably varied and comprehensive assemblage of commonplaces and truisms involving the relevant
problematic terminology. A well-known problem for such an approach is the problem of “epiphenomenal” truisms—things commonly asserted as truistic by the folk, but which seem to play little practical role in guiding the application of the relevant terms in contexts other than “metalevel” reflexive examination of the terms themselves. This is likely to stem, for instance, from the incorporation of bits of religion, pop philosophy, or pop science into ordinary discourse. For example, given the prevalence of popular dualist views, it is almost inevitable that analytic functionalists who are materialists will need to dismiss some dualistic tendencies in explicit “folk psychological” theorizing on some such grounds.

An alternative is to pay more attention to actual linguistic conceptual practice. When the folk are not asked to make explicit what they believe to be analytically or truistically the case about the problematic domain, but rather observed with an eye to how in practice relevant terms are applied, withheld, and apparently conceptually linked, what theory can reasonably be seen as implicit in that practice?

As described the task is potentially massive, fit for an army of ethno- and psycholinguists. Philosophers have a couple of ways of cutting it down to size. One consists in deploying a philosophical “nose” for those aspects of ordinary usage that are potentially especially revelatory of those aspects of implicit folk theory that bear on the major points of philosophical controversy about the problematic domain. For instance, in debates about analytic functionalism about the mental, little time is spent exploring the functional distinctions if any between anger and rage, boredom and ennui, frustration and disappointment. We know we can let such issues lie where they fall without significantly affecting the central ontological issues. The real task after all is not, typically, to articulate a complete folk theory, so much as to refute philosophical objections to the possibility of doing so in a way that will render it consistent with particular metaphysical claims.

This resource of selective attention is typically incorporated into the “method of cases.” Rather than observing speakers’ actual usage in real-life circumstances, one describes actual or possible situations—chosen for their capacity to isolate and clarify points of philosophical contention—and elicits intuitive judgments about what is or would be the correct thing to say in, or about, those situations. ‘Intuitive’ judgments here means, roughly, judgments that arise spontaneously as manifestations of a subject’s ordinary, tacit linguistic knowledge, rather than being deduced from their explicitly held, amateur or professional philosophical, psychological, or linguistic theories.
The method of cases has the virtue that a philosopher can claim *qua* competent speaker of ordinary language to be as good a representative of what it is appropriate to say about such cases as any other: the relevant “folk theory” is conceived as, near enough, implicit in the shared linguistic competence of normally competent members of a given speech community. But it is not surprising that there is less than complete agreement between philosophers regarding what should be said about actual or possible test cases that bear on live philosophical controversies. The claim that one’s intuitions are completely unaffected by one’s philosophical views or background is as implausible as the claim that the intuitions of the folk generally are completely unaffected by religion, superstition, pop psychology, and the like, and once one turns from observing actual unselfconscious practice to consulting intuitions under cross-examination, the risk must increase that answers are affected, even if unconsciously, by a theory that is “epiphenomenal” with respect to practice (see Johnston 1987 for a useful discussion of this problem).

Also crucial in what follows is the point that any such method will be hostage to the assumption that, for relevant philosophical purposes, the folk speak as one. If philosophers fail to agree then we can turn to such paradigm masters of unadulterated folk terms and concepts as passengers on the Swanston Street tram or the Bondi bus. But if even those representatives of the folk turn out inconsistent with one another at crucial points, the method of cases will still be in trouble.

There are of course various resources for “explaining away” some such discrepancies. One may hope to argue that there is a common core of agreement that demarcates a shared concept. Discrepant judgments are to be considered “peripheral” rather than “core,” and at least some of those who make these judgments will be considered to be in error relative to some truth determined by the core, conceived as relatively a priori, combined with relevant facts. In one sort of case, the core amounts to a canonical definition of a kind of theoretical entity, by way of a definitive functional role it plays relative to the uncontested domain; the periphery of contested claims consists of true or false a posteriori claims about that theoretical category. In another sort of case, the core will be thought of as a set of definitive principles, and the periphery will be thought of as consisting of claims held, in some cases mistakenly, to be deducible from those principles—claims, in other words, resulting from attempts, in some cases mistaken, to apply those general principles to particular cases. (In cases of both these kinds, the idea is that a “correct” account of the peripheral cases in one way or another follows from the core of agreement, together with
further matters, not of opinion, but of fact.) As well, or instead, there is the option of discerning or positing a relativistic or indexical element in the concepts in question: as portrayed in rather stark simplicity in the “Vegemite” dialogue.

Each of these possibilities will be salient at points in what follows. For the moment let us note that in any application of the Canberra Plan, the relevant folk theory must pretty well inevitably be thought of as some cleaned-up, systematized abstraction from an extremely complex host of facts about assorted points of usage and intuition. Ramsification proceeds by way of generating a schema (an “open sentence”) from a theory by substituting variables of quantification for relevant theoretical terms in that theory. Respective members of the unique tuple of entities, if there is one, which (“near enough”) satisfies that schema, are thus identified as the referents of those theoretical terms. Since no tuple has inconsistent properties, the price of attributing an inconsistent theory to the folk will be arriving at an “error theory” of the problematic domain. Often it will seem more appropriate to remove inconsistencies from the folk theory than to construe the folk as, literally, talking about nothing. For similar reasons, it will sometimes seem more appropriate to drop some claims from the folk theory because they simply clash with well-established empirical findings. Thus a “revisionist” element may be hard to avoid, in the attempt to construct a folk theory that is to have any prospect of identifying the subject matter of folk discourse. At the same time, the point of the exercise will be lost if the folk cannot plausibly be held to be in some sense, however implicitly, committed to the relevant folk theory. The latter point, I suggest, should count as an important constraint on Canberra Plan methodology.

4 Can Moral Functionalism Avoid Relativism?

Focusing our attention, then, on the case of ethics, we can hardly avoid noting that moral disagreement is pervasive and robust. Moreover, people often give firmly held moral opinions “bedrock” status, refusing to be argued out of them by appeal to shared common principles (let alone by explicit appeal to some domain of moral theoretical entities such as “moral facts” or “moral properties” conceived as somehow causally or otherwise bestowing some or other moral status on the disputed cases). Note too that this is far from a matter of giving mere theoretical lip service to the relevant points of view: such opinions are frequently acted on in the face of considerable obstacles. Taking not the mere fact of moral disagreement, but
these kinds of recalcitrance, into account, different parties can appear, if
using moral terms to predicate descriptive properties at all, to be using
them to predicate different descriptive properties. Moral functionalism, in
other words, can threaten to lead to moral relativism.

One kind of tactic to avoid this kind of consequence—one that Jackson
invokes—is to conceive of ourselves as Ramsifying a suitable future or
counterfactual substitute for actual, present folk morality. Jackson calls this
“mature folk morality.” He says:

We can think of this story (John Rawls’s) as one story about how folk morality
should evolve over time: we modify folk morality under the constraint of reconciling
the most compelling general principles with particular judgments. In this way we
hope to end up with some kind of consensus.

In any case, however we should characterize the way folk morality is evolving
over time, it is useful to have a term for where folk morality will end up after it has
been exposed to debate and critical reflection (or would end up, should we keep at
it consistently and not become extinct too soon). I will call where folk morality will
end up, mature folk morality. The idea is that mature folk morality is the best we
will do by way of making good sense of the raft of sometimes conflicting intuitions
about particular cases and general principles that make up current folk morality.
(Jackson 1998a, 133)

This kind of tactic comes in several varieties, and Jackson, in the above
passage, is not always clear which he advocates. The simplest version,
which he does not advocate, simply assumes that there will be a time when
moral debate and critical reflection have run their course. Moral disagree-
ment not due to simple errors of fact or reasoning will be done and gone,
and everyone not subject to such errors will subscribe to mature folk moral-
ity. The moral properties are those that would be picked out by way of
Ramsifying that theory.

Such a view is already one step away from a view that takes people’s
actual, current, moral concepts as a guide to what they are talking about in
moral discourse (unless there is actual current unanimity among the folk—
or evidence in their practice—that the content of actual future folk moral-
ity better expresses the content of their current moral concepts than their
current moral views). But as Jackson reveals, we must also allow for the
possibility that fundamental moral debate will last as long as the human
race. For this reason if no other, Jackson must envisage a potentially merely
counterfactual “mature folk morality”—what we will or would wind up
with. Something further emerges in the requirement that we “keep at it
consistently,” and in the mention of Rawls’s view as bearing on how folk
morality should evolve over time: the risk of a “bogus” folk morality being
what actually will or would emerge at the end of fundamental moral
debate. If this is a coherent possibility, then, arguably, it is not how folk
morality actually will or would, but how it should (given the chance) evolve
over time, that really matters. And in that case moral functionalism not
only needs a substantive account of the constraints on moral debate and
reasoning required in principle to distinguish genuine from bogus mature
folk morality: it must be reasonable to judge that the folk are sufficiently
committed, explicitly or implicitly, to those constraints for it to be rea-
sonable to construe whatever version of mature folk morality those con-
straints would mandate, as taking priority over their current fundamental
moral commitments, in best expressing their actual, current moral
concepts.

Before defending the claim that moral functionalism must meet such a
challenge, we need to note the fact that Jackson does endorse the fall-back
possibility of a relativist moral functionalism. It is worth quoting an exten-
sive passage here:

I have spoken as if there will be, at the end of the day, some sort of convergence
in moral opinion in the sense that mature folk morality will be a single network of
input, output, and internal role clauses accepted by the community as a whole. In
this case we can talk simply of mature folk morality without further qualification.
Indeed, I take it that it is part of current folk morality that convergence will or would
occur. We have some kind of commitment to the idea that moral disagreements
can be resolved by sufficient critical reflection—which is why we bother to engage
in current folk morality. But this may turn out to be, as a matter of fact, false. Indeed,
some hold that we know enough as of now about moral disagreement to know that
convergence will (would) not occur. In this case, there will not be a single mature
folk morality but rather different mature folk moralities for different groups in the
community; and, to the extent that they differ, the adherents of the different mature
folk moralities will mean something different by the moral vocabulary. . . . I set this
complication aside in what follows. I will assume what I hope and believe is the
truth of the matter, namely, that there will (would) be convergence. But if this is a
mistake, what I say in what follows should be read as having implicit relativization
clauses built into it. The identifications of the ethical properties should all be read
as accounts, not of rightness simpliciter, but of rightness for this, that, or the other
moral community, where what defines a moral community is that it is a group of
people who would converge on a single mature folk morality starting from current
folk morality. (Jackson 1998a, 137)

Several things need noting in this passage. One is how little attention is
paid to the question of what would count as the right kind of critical reflec-
tion, or the right kind of resolution. It is all very well to say that we share
some kind of commitment to the idea that moral disagreements can be
resolved by sufficient critical reflection, but that provides no basis for saying our ethical concepts are in essence held in common, if we have or would have differing conceptions of what would count as suitable “critical reflection.”

Second, it is to be noted how Jackson slides into speaking of what the adherents of the different mature folk moralities will mean. If we wish to give an account of the subject matter of contemporary moral debate and discourse, the issue should not be what they mean, but what we mean: and every counterfactual antecedent, every distinction between their circumstances and ours, draws a prima facie wedge between the two.

Third, it’s worth noting that the more moral questions left unresolved, and the less correlation between people’s individual answers to each of them, at the end of moral debate, the smaller and more numerous the “moral communities” will become. If we think of a continuum between a radically particularist view of moral judgments, and an opposed extreme that holds that a very few very general moral rules suffice in principle to settle all moral issues, then Jackson, as an optimist about moral convergence, seems to be judging that the truth—at least as it would be revealed by suitable critical debate—lies nearer the latter end of the spectrum. But the less that is true, the worse, I suggest, his problems get.

Finally, and most importantly for our purposes, Jackson says nothing directly about the residual “disagreements” that might remain between distinct “moral communities.” But a straightforward implication of what he says is that they will simply amount to “failures to disagree.” Members of different communities will mean different things by their ethical terms, critical reflection will have removed all factual errors and mistakes in reasoning, and the residue will be a number of communities within which there is fundamentally perfect ethical agreement, and between whom there is mere “talking past one another.”

In short, Jackson needs some help from quasi-relativism, if he is not (by my lights) to misrepresent this situation, given that he cannot rule it out. In fact, it’s rather worse than that, if we imagine, as Jackson must, somehow bridging the prima facie gap between what such future or hypothetical communities would be talking about, and what we and our contemporaries are currently, actually talking about when we use the moral vocabulary. I now argue (roughly speaking) by dilemma. (There are further options we could consider: but plausibly they are mere intermediate cases that don’t improve the available outcomes.)

First option: for each current user of moral vocabulary there is a determinate “moral community” they would, under suitable counterfactual
assumptions, belong to such that the meanings of terms in their moral vocabulary should be seen as the same as the meanings of terms in the moral vocabulary of that moral community. *Bad consequence:* there are two types of apparent moral disagreement. There are moral disagreements between people affiliated with the same (hypothetical) moral communities, and these are genuine disagreements—disagreements, so to speak, that reflect disagreement or ignorance about which judgment the relevant moral community would make. Other apparent moral disagreements are really nondisagreements, since the disputants are affiliated with distinct moral communities; hence, meaning what members of those communities would mean, they are not genuinely in disagreement, merely at cross-purposes. *Worse consequence:* since none of us knows what comprises the full set of (future or counterfactual) moral communities, let alone which of us is affiliated with which community, none of us knows in the case of any particular apparent moral disagreement, whether it is by these lights real or bogus.

*Second option:* typically, we contemporary users of moral vocabulary are not, as individuals, determinately associated with particular future moral communities. In particular, there is no entailment from what members of any particular one of those moral communities will or would mean by terms in their ethical vocabulary, to what we mean by those terms in our vocabulary. *Bad consequence:* We either mean something entirely different, determined by our actual current moral judgments—vitiating the role of mature folk morality—or at best we mean something vague that is indeterminate with respect to matters over which those different communities differ. The latter alternative is unattractive precisely because the points over which those hypothetical moral communities differ are likely to include matters over which contemporary users of the ethical vocabulary are in fundamental, bedrock disagreement. If such bedrock disagreements suffice to make it true that the hypothetical communities differ in what they mean, it seems odd that they should not have similar consequences for those involved in contemporary moral debates.

Jackson does not really expect that this relativist outcome will or would come to pass. But if it merits even a modicum of credence, his view seems to leave a lot to be desired so far as addressing the topic of moral disagreement. Jackson’s proposals raise a panoply of highly debatable questions about what the outcome of critical debate must be, how it may be arrived at, and what connection it has, exactly, with the semantic knowledge and intentions of contemporary folk. And if we take current speakers’ actual moral judgments (as displayed both in speech and in action) as revelatory
of their ethical concepts, we are in danger of having to describe as non-disagreements disputes that stubbornly resist such a characterization.

Interestingly Jackson makes one remark that seems to run counter to what ought to be the underlying assumption of his proposal of mature folk morality, namely, that there is a bridge between our current moral talk and the meanings of moral terms as they would figure in mature folk morality, such that the latter determine the common subject matter of our contemporary moral debates. He says:

The principles of folk morality are what we appeal to when we debate moral questions. They are the tenets we regard as settling our moral debates. ‘... It would be a betrayal of friendship not to testify on Jones’s behalf, so I’ll testify.’... The dispute-settling nature of such a tenet shows that at the time in question and relative to the audience with whom we are debating, the tenet is part of our folk morality. If there were not such benchmarks we could not hold a sensible moral discussion with our fellows. Nevertheless these benchmark tenets are far from immutable....

What is, though, true is that there is a considerable measure of agreement about the general principles broadly stated. (Jackson 1998a, 131–132)

Apart from the optimistic focus on what happens when moral debates do get settled, Jackson here goes directly against the idea that the essential common ground in our debates is a commitment in principle to the deliverances of mature folk morality, whatever they may be. For he insists that it is currently endorsed but mutable tenets that provide the common ground that is essential if there are to be sensible debates. But in the light of these comments it is hard to see how mature folk morality can lay claim to settle the subject matter of our current debates. As for the “general principles broadly stated,” insofar as there is robust disagreement about particular cases, it has to be doubted whether those who subscribe to such principles thereby interpret and apply them in such a way as to rule out the possibility that they are applying moral concepts that differ importantly in their application to those particular cases.10

5 Basic Logical Disagreement Despite Differing Application-Conditions?

Before going any further we need to dispose of an important and pressing objection to the whole way I have laid out the issues so far. Here’s how it goes. In the above cases, the failure of basic logical disagreement, which makes the claim that there is “real” disagreement problematic, stems from differences in extension between words or concepts employed by the disputants, due to differing application-conditions for those words or
concepts. But often the extensions of terms or concepts are not best seen as settled by associated application conditions, and then such situations don’t arise. If that’s how it is with all putative ethical quasi-disagreements, then there is no call for a quasi-relativist position. Let’s look at some cases of that kind.

Water dialogue

**Bargle:** Surely it’s a conceptual truth that water repels fire? Where there’s water, there’s no fire: isn’t that why we put out fires with water?

**Argle:** Your water-concept is not merely mistaken, it’s positively mediaeval! Don’t you know that a burning reel of nitrate film will continue to burn, even underwater, since nitrate film generates oxygen as it burns? Magnesium also burns underwater!

### 5.1 Anchoring and Bypassing

Here Bargle’s concept is bypassed: he is talking not of some mythical substance that repels fire, but of the natural kind called ‘water’. Despite the difference between current and previous theories of water, we can refer to the same stuff as those who held those contrary earlier theories, because of the existence of that natural kind to which, we reasonably say, it was always intended to refer. The kind anchors uses of the term, enabling commonality of reference even when people’s water-concepts are idiosyncratic, mistaken, or out of date. This is just like the familiar case of theory change with persistence of reference, as in evolving theories of the electron. Bargle’s idiosyncratic concept is irrelevant and there is no threat of failing to disagree. Notice here that the circumstance enabling the concept to be bypassed also blocks the “no-fault” condition.

But what if water-as-natural-kind itself turned out mythical, with no real natural kind anchoring uses of the term? Insisting the term must refer to a natural kind or nothing would leave debates about water and water-concepts resembling debates about unicorns and unicorn-concepts. Quite likely we would revise our ideas about the semantics of ‘water’, and go on using the term for various substances more-or-less occupying something resembling the traditional water-role. But that would make it vague, so that people might sometimes legitimately say ‘you may call this water, but I don’t’, just as they sometimes say ‘you may call that color green, but I don’t’. There are preconditions for “anchoring” a term without which a default assumption or even a stipulation, that all within the speech community shall use that term with common reference, will be undermined.
5.2 Social Kinds and Anchoring, Perfect and Imperfect

Not that a term has to be or even resemble a natural kind term to be reasonably firmly “anchored” and so “bypass” any associated concept. Consider:

Second football example

Argle: In football, it is permissible to pass the ball by punching it.

Bargle: It is not!

Argle: Oh, you must be thinking of soccer! I’m talking about Aussie rules.

Bargle: And so am I! I’ve known the game since childhood and my concept of it is very clear. Passing the ball by punching it is not permitted.

We must not conflate dispute over concepts with dispute over things those concepts apply to. This case involves genuine disagreement over concepts of Australian Football: but is it also a genuine dispute about games the rival concepts apply to?

The question is ambiguous. The concepts invoked have different application conditions, hence different extensions (‘it’s permissible to pass the ball by punching’ being a necessary condition for a game’s being in the extension of Argle’s concept, but not Bargle’s). In this sense of ‘apply’ there is no misapplication of concepts; thus we might say each is right in his own terms, and this is another case of failure to disagree.

But that would be wrong. We should more defensibly say Bargle is wrong and his concept (given his words) is mistaken. This is because we take Argle and Bargle to succeed in commonly referring to and disagreeing over a single well-defined football code, about which Bargle’s claim is false. Bargle uses the words ‘Aussie Rules’, in a context which by default licenses assigning to the parties’ words whatever reference they normally have in the relevant speech community. This gives another sense (though perhaps a stretched one) of ‘what the concepts apply to’, allowing a concept to be misapplied. Because an idiosyncratic concept like Bargle’s can be “bypassed,” Bargle achieves reference despite it, to the same code (and games) of football as Argle. Hence we have no quasi-disagreement here (in the sense of ‘quasi-disagreement’ introduced earlier).

However, we must remember “Wittgenstein’s lemma”: not all games are like Aussie rules. Many games are not so clearly codified. More generally, there are many social kinds, games among them, such that our names for them are vague in respect of many details one might expect to be
determined a priori by the relevant concepts: and thereby, disputes may arise. Think for instance of ‘common law marriage’, and ‘gay marriage’. Are these “real” marriages? Opinions differ, and debates about the matter often take the form of debates about what ‘marriage’ really means.

Cultural variations in marriage customs complicate the semantic issues. When English-speakers speak of ‘marriage’, they don’t speak of it only as it exists in their own English-speaking community. A colleague recently witnessed in Northern China a wedding ceremony in which two brothers married a single woman. In the preceding (true) sentence I apply an English-speaker’s concept of marriage to a form of that institution that is not part of a typical English-speaking community, nor of any of the atypical English-speaking communities to which I belong.

In any case, there are many terms that allow limited variation in reference, tracking variation in concept. ‘Sofa’ appears to be one. Usage varies, so argie-bargie easily ensues—and in that case, as in many others, is almost as easily resolved.

The distinction between natural-kind terms and others is sharpish, but nonnatural, especially social, kinds form a continuum, from cases like ‘Australian Football’, via ‘hopscotch’ and ‘sofa’, to cases like ‘marriage’ or ‘chief’. The further along this spectrum, the less “anchoring” is available to stabilize reference across time and speech-community, so the more room and need for individuals’ concepts to play a complementary, disambiguating role in reference determination.

Thinking about translation helps to illustrate these issues. Consider the following dialogue:

*Anthropologists’ dialogue*

**Anthropo-Argle:** The two most important figures in the socio-politico-religious hierarchy of the Sregnabløy people are the namriahc and the reganam. For our purposes we can best translate those terms as ‘king’ and ‘prime minister’, respectively.

**Anthropo-Bargle:** All wrong as usual! The one you call ‘king’ is the high priest, and the one you call ‘prime minister’ is the king.

Here either choice of labels might be permissible, but one choice might fit Argle’s usage better, the other Bargle’s. Though Anglophones will agree on many paradigms, there may be insufficient anchoring of the terms for either disputant to be clearly misusing them in a case such as this. When we go from culture to culture trying to identify moral concepts, we are similarly likely to find ourselves at times inevitably wavering. Is this a very strange morality we encounter here? Or a comparably strange
religion? Or an odd political tradition? Or some bizarrely overblown commitment to notions of “honor,” or of “etiquette”? However clear the answers may be in some cases, in others there may simply be no clear answer.

I’ve suggested that for some terms, reference is “anchored” for a speech community, “bypassing” concepts individuals take as analytically expressive of their meaning: mechanisms for this can implement the default convention that terms shall be used with a common reference within that community. Sufficiently robust anchoring can block “no-fault” disagreements of the kind we’re considering. For other terms, prerequisites for such “anchoring” mechanisms are missing or incomplete. Then an individual’s concept cannot be wholly “bypassed,” since it is required to substitute for, or complement, those mechanisms: these cases allow for “no-fault” disagreements.12

6 Contestable Normative Notions in Jackson’s Account—”Critical Reflection” and “Rationality”

We can now make a further comment on the problem Jackson faces in “bridging” the prima facie gap between current and hypothetical future moral concepts, seen as implicit in current and “mature” folk moralities. The Water dialogue and the Second Football dialogue remind us that (in old-fashioned terminology) differences or changes in concepts or “sense” need not entail discrepancies or variations in reference or extension. But this is possible only to the extent that there exist determinate relations of natural-kind similarity and difference—or clear and canonical social conventions—capable of trumping or replacing concepts in determining or “anchoring” reference or extension, in a way that keeps reference or extension constant across the relevant differences or changes. Jackson’s extensive contributions to the theory of reference notwithstanding, I claim that we do not have (and cannot find in From Metaphysics to Ethics) what would be needed to tell us that moral terms in contemporary use, and their descendants that would figure in mature folk morality, are “anchored” in such a way as to refer respectively to the same moral properties. There are various remarks about our shared implicit commitment to accepting the results of critical reflection, coupled with descriptions of mature folk morality as the future or hypothetical product of critical reflection, but in the light of the foregoing discussion, it seems to me that these simply do not establish the required commonality of anchoring. Jackson himself in fact says: “critical reflection is, by definition, what any theory should be
subjected to. Of course, precisely what critical reflection on current folk morality comes to in detail is a matter of considerable debate” (Jackson 1998a, 139). He says also: “What is a priori according to moral functionalism is not that rightness is such-and-such a descriptive property, but rather that A is right if and only if A has whatever property it is that plays the rightness role in mature folk morality, and it is an *a posteriori* matter what that property is” (ibid., 150–151).

But it is not clearly a priori that this commitment to the canonical status of mature folk morality can be given the requisite degree of determinacy while itself surviving critical reflection, given the possibility of debate about the proper kinds, methods, and circumstances of critical reflection, about which individuals, and in what psychological states, are appropriate participants, and so on, and consequent vagueness in the idea of mature folk morality. Nor is it clear that it is a commitment capable of trumping current other commitments, in giving an account of what moral terms *currently* mean. Compare: I may preface a book with the words ‘I am sure some claims in this book are false’. In no way does this abnegate my current commitment to any individual claim in the book. I cannot get “off the hook” for false opinion by saying ‘it was always my intention to tell the truth’.

There are two strands of discussion that raise somewhat similar issues in Jackson’s ethics chapters—chapters 5 and 6—in *From Metaphysics to Ethics*. In principle they are distinct but at the end of the day they are closely interrelated, and for me, similar doubts arise about them. First—mostly in chapter 5, “The Location Problem for Ethics”—there is a defense of descriptivism and the promotion of “moral functionalism” as the right way to implement it. Jackson stresses the “ecumenical nature” of moral functionalism, and mentions a number of different lines along which mature folk morality might develop, maintaining on the whole an agnostic stance as to which of those possible outcomes will, would, or should, come to pass. My critical comments so far have essentially been directed to the generic account there given of “mature folk morality,” and the role assigned it.

But in chapter 6, “Analytical Descriptivism,” and especially in the closing sections of the chapter, where Jackson deals with the Open Question argument, and the connection between moral judgment and motivation, he needs to put a few more of his own cards on the table, and express some opinions about some of the content that mature folk morality will have. (Actually, some of these commitments already surface in chapter 5.) He continues to allow for a range of options, but they are all variations on a theme. The theme is that a correct account of the subject matter of moral
judgment must be a “response-dependent” one. This notion is first mentioned as an attractive option, early in chapter 5: “Thesis (S) [the global supervenience of the ethical on the descriptive] is compatible with the idea that ethical nature . . . is in part determined by facts about our responses and attitudes, with the appealing idea that, in Mark Johnston’s terminology, value is response-dependent” (Jackson 1998a, 120).

There are further remarks in chapter 5, which reveal Jackson’s approval for this option, but at this point let us note an important passage from chapter 6. He says:

Moral functionalism sees the meanings of the moral terms as given by their place in a network. Part of that network is certain output clauses that tell us how beliefs about ethical properties connect with facts about motivation. The details of these output clauses are highly controversial. But, to fix the discussion, let’s suppose that the connection with motivation goes roughly as follows . . . it seems to me to be on essentially right lines . . . (a) right act is one that has properties of value to an extent that exceeds that of the various alternatives to it, and a property’s value depends on its being rational for us to desire it. The moral rightness of action is, then, a matter of its having properties of value . . . that pertain to morality. . . . It is important to this sort of proposal that we can give an account of what makes it rational to desire a property that does not reduce, uselessly, to its being a rightness-making property. (Ibid., 156)

After listing some recent proposals for identifying properties it is rational to desire, Jackson continues:

The details will not matter for what follows. What will matter, though, is something I take to be widely agreed, perhaps under the heading of the rejection of Platonism about value, perhaps under the heading of the response-dependence of value. It is that what confers value on a property ultimately comes down to facts about desires. . . . Accordingly, this much is right in subjectivism about value: what gives value whatever objectivity it has comes down, somehow or other, to some combination of facts about the convergence, the stability, the coherence between first-order and higher-order desires, the desires of idealizations of ourselves, the desires of our community, and the like. (Ibid., 157)

A central tenet of my own view is that there will always be the possibility of bedrock disagreements about what counts in this sort of context as rational. This is inevitable given persisting bedrock moral disagreements, one ethicist’s modus ponens being another’s modus tollens. This is why I claim that there may be moral quasi-disagreements in my strong sense: irresolvable disputes in which there is no uncontrovertially “neutral” standard by which at least one party may be judged to be mistaken. We have no better standard than ideally rational judgment under ideal circum-
stances: but the notion of “ideally rational” judgment lacks sufficiently determinate anchoring to eliminate bedrock disagreements, in a similar way, and for similar reasons, to moral concepts themselves. This is a main reason for preferring ethical quasi-relativism. And in his rejection of Platonism, his critique of “Cornell realism,” and his other doctrines, Jackson makes clear that he has cut himself adrift from other potential dispute-settling anchors for moral concepts.13

These two strands of Jackson’s argument—the general idea of mature folk morality as implicitly specifying what the moral properties are of which we speak, and the particular adherence to a response-dependent account of value, revolving around the idea of rationality in desire—are in principle distinct. But in practice it appears to me that these operate as two sides of a single coin, in the view Jackson is outlining, and some similar criticisms apply to both. What counts as ideal “rationality” in desire, and its import for the identification of moral properties, and what counts as the right kind and context of critical reflection, and the general role of the mature folk morality it will, would, or should arrive at, are all matters that must be hammered out in arriving at mature folk morality, and the determination of these different matters will be part of a single process. Even setting aside the “bridging” issue, these interwoven processes cannot do their jobs unless subject to appropriate norms, and the content of those norms must at the end of the day itself be a matter for quasi-disagreement, not subject to resolution from any uncontestedly “neutral” standpoint.

7 Words versus Concepts

Thus my view is, roughly, that for words like ‘right’ and ‘good’ (in the moral sense)—and I could add, ‘person’—we can say that “anchoring” to some degree fails, leaving room and need for individuals’ concepts to play a role in reference determination, and no impartial way to adjudicate between speakers so as to secure a perfect match. I say ‘roughly’ partly because this formulation suggests that the view is more strongly tied than I intend to claims about particular linguistic expressions. So far I’ve mentioned the application conditions of words, and of concepts, more or less as if these are interchangeable ways of speaking. Now I need to put in a correction. Compare Jackson:

I use the word ‘concept’ partly . . . to emphasize that though our subject is the elucidation of the various situations covered by bits of language according to one or another language-user, or by the folk in general, it is divorced from considerations local to any particular language. When we ask English users in English for their
intuitive responses to whether certain cases are or are not cases of knowledge, we get information (fallible information . . .) about the cases they do and do not count as covered by the English word ‘knowledge’. But our focus is on getting clear about the cases covered rather than on what does the covering, the word *per se*. We mark this by talking of conceptual analysis rather than word or sentence analysis. (Jackson 1998a, 33–34)

This is an unsatisfactory and ambiguous passage. Just what does “getting clear about the cases covered” amount to? We might think of conceptual analysis as combining a part where suitably varied possible situations are divided into those ordinary English speakers do and those they don’t count as (in this example) “knowledge”—here an English word is definitely central to the inquiry—and another part attempting to discern principles or patterns governing the resulting division. The latter part makes no reference to language as such, but that hardly “divorces” the whole process from considerations local to English, whatever we say “our focus” is on. Perhaps no other language uses a word that divides cases in just the way ‘knowledge’ does? Epistemologists might, after all, investigate truth, justification, belief, reliability of belief-generating processes, flukes, hidden defeaters, and all the rest of it, without logically compounding them in the precise manner associated with that English word: there is no a priori guarantee that any language but English does that, and the more convoluted the twists and turns of Gettierology, the more likely it may seem to be an enterprise entirely local to English.

The issue of “anchoring” is with us again. If conceptual analysis of an English word like ‘water’ points to its naming a natural kind, we can happily turn the job of further investigation over to the chemists, thereby pushing away the ladder of English language, and considerations “local” to it. But when we speak of abstractions—knowledge, virtue, causality, probability—we cannot anchor our conversation by planting the relevant thing on the table or decanting the relevant stuff into a test tube, saying ‘*This* is what we’re talking about’. We must, rather, use some language or other to identify our subject matter. Whether this makes our inquiry permanently parochial to that language is a question to be answered on a case by case basis. In principle, the answer should in part depend on how translation into other tongues fares. We may have some shrewd hunches based on our knowledge of human nature and other cultures, but a hunch is a starting point for inquiry, nothing more.

Consider another of Jackson’s remarks: “I said, following Humpty Dumpty, that we can mean what we like by our words. But if we want to speak to the concerns of our fellows, we had better mean by our words what they mean.
If we are interested in which property the word ‘right’ in the mouths of the folk picks out, we had better give a central place to folk opinion on the subject” (Jackson 1998a, 129). Ethical quasi-relativism takes seriously the possibility that there is no one such property as the property that is picked out by ‘right’ in the mouths of the folk. But regardless, if we wish our inquiry into folk opinion on ethical matters to transcend a local concern with English, we will need to arm ourselves with something more than a resolve not to gratuitously deviate, Humpty-Dumpty-wise, from ordinary usage of the English word ‘right’. Whichever words we use, if we are to find a way to communicate with the polyglot folk about ethical matters, we need to find or create linguistic means, simple or complex, for discussing ethical concepts with them. Roughly speaking, we need to come up with a suitable translation of the word ‘right’ into their language. So we need, roughly, to figure out how they express a concept the same as (or sufficiently similar to) the one we are expressing: then—but perhaps not till then—we may have some confidence of investigating considerations “not local to any particular language.”

(As Quine has taught us, issues about translation usually have analogues within a single language community. Mark Twain well understood this. When Huck Finn said “I got aboard the raft, feeling bad and low, because I knowed very well I had done wrong, and I see it warn’t no use for me to try to learn to do right,” he was in many ways giving a central place to folk opinion on the application conditions of ‘right’ and ‘wrong’: but we know that he did what he felt was right, though neither he nor members of his linguistic community called it that. Thus his word ‘right’ did not express his concept of what was right. It’s important to note that, and how, we can know this.)

For all the vagueness of concept-talk, then, I think we do better to accept a practical need to talk of concepts—and to find good methods of analyzing and individuating them—in ways that do not easily reduce to a coy way of talking about particular words and word-meanings.

Generally, given due allowance for two-dimensionality, semantic externalism, and such, I take it to be perfectly appropriate to talk of ‘the concept expressed by the English (or French, Chinese, or Etruscan) word w’. But only a subset of concepts corresponds neatly to word meanings, and many important ones do not: I include ‘identity’, ‘person’, and ‘right’ among them. The words I just used are English words (or philosophers’ variants), but those concepts could in fact figure centrally in the lives of people whose language had no such words, and I maintain that analyzing them need not be best understood as doing linguistic analysis of the words associated with them.16 People’s person-concepts, for instance, are mostly not
deployed by using words like ‘person’. It would be as misguided to take analysis of that word’s meaning as a guide to the structure of those concepts, as to question Leibniz’s law by citing usage of the word ‘identical’.

8 Modeling Quasi-disagreement

How can the quasi-relativist best represent the “kind of” disagreement involved in “quasi-disagreements”? For this I borrow from Michael Dummett’s work on concepts.

Dummett is well known for suggesting that logical constants may be characterized either by natural deduction introduction rules, or elimination rules, and that these rules should be subject to a mutual “harmony” constraint, so that adding a logical constant to a theory results only in a conservative extension of that theory (e.g., in Dummett 1991, 210–218). Dummett also, in his discussions of truth and assertion, draws parallels with the concept ‘a win in chess’. One ignorant of competitive games would need to learn that the goal is winning, not merely which positions count as “wins” (and for which player), in order to acquire this concept. Dummett (1981, 296–297) makes analogous claims about the difference between merely knowing sentences’ truth conditions, and knowing how to make assertions using them. These are all instances of a general distinction he draws between conditions for a concept to apply, and the consequences of applying it: he holds that each of these distinct aspects is to be considered constitutive of a given concept. This allows a richer and more nuanced account of aspects of concept-individuation than simple application conditions accounts can provide. For present purposes, the consequences to be taken into account may be in our theories, in our discourse generally, or be practical consequences in our lives, including specially, our evaluative and affective lives.

We could see this distinction as being between, broadly speaking, two (very different) aspects of ‘content’. It may not be applicable to every concept, but it seems to me useful for making sense of quasi-disagreement. Let’s stipulatively label the set of those consequences of applying a concept that are partly constitutive (or individuative) of it, the C-role of that concept. This is to distinguish the “role” in question, both from the concept’s application conditions (which could also be called part of its “role”), and from the functional roles specified by (for instance) folk psychology and folk morality. I emphasize yet again: we are not talking merely of inferential consequences, but of consequences in general, including practical consequences—attitudes, emotions, and actions.
The point is not that consequences that are constitutive or individuative of a concept should always follow from its application: but if those consequences never followed, for an individual or a population, that individual or population could not be said to possess or apply that concept. (This mirrors the thought that an extreme psychopath might be said to lack the concepts of right and wrong, even if he or she could apply the words ‘right’ and ‘wrong’ in a manner that exactly mimicked the usage of someone highly moral. We would not on the other hand be inclined to describe people as temporarily ceasing to possess those concepts every time they fall into episodes of apathetic or narcissistic amorality.)

The idea is that we can view some concepts as having some kind of commonality—as suitably “the same”—despite having different application conditions. Sometimes, I think, we in effect use ‘same concept’ so as to give priority to application conditions, sometimes so as to give priority to a concept’s C-role. Generally (though it’s context dependent) the more evaluative, and especially the more normative, a concept, the more we prioritize C-role in concept individuation; the more descriptive, the more we prioritize application conditions. But the general idea is that in quasi-disagreements, the relevant concepts differ in their application conditions, but are competing to play the same C-role. The difference in application conditions accounts for the lack of basic logical disagreement, but competition for the same C-role—the clash of competing criteria for the same consequences—accounts for there being, nonetheless, a kind of “real” disagreement. In cases that interest us this may be not theoretical (or “basic logical”) disagreement but practical disagreement, in a social context: this is why there often needs to be competition rather than peaceful coexistence between moral concepts.

Commonality of C-role can also, of course, account for other aspects of apparently successful communication between speakers who use the same word to express concepts with different application conditions. Indeed, it provides an extra resource for dealing with the difficulty I raised for Jackson, of establishing a plausible bridge between current actual and hypothetical “mature” folk morality, capable of underwriting the claim that in spite of the differences, in some important respect terms in those two theories “mean the same.” Not that this provides any solution Jackson can use, since it in no way establishes that the same properties are referred to by the relevant terms of the two theories.

Likewise, we are now in a better position to understand the enterprise of translating the moral language of another culture with whom we are in moral disagreement. Difference in application conditions need not obstruct
us in translating a word of theirs into English ‘right’, if it is clear that it plays something like the same C-role for them that ‘right’ plays for us. Not that identity of C-role alone should suffice to identify a word of theirs as to be translated by our ‘right’. Too large a gulf between the application conditions for their word, and for ours, would rule out such a translation, just as would a gulf between the respective C-roles. Still, approximate commonality of C-role should carry substantial weight in identifying suitable translations. Conversely, a good way to get a grip on what I am here calling the relevant “C-role” is by reflecting on what we would give weight to in such a translation situation.

It would be misleading, however tempting, to say that arming ourselves with these borrowings from Dummett will enable us to individuate concepts with increased precision. All this talk of “concepts,” of extracting a coherent folk morality from the intuitive judgments of the folk, and so forth, concerns quite heroic abstractions from an extraordinarily complex host of events involving individual utterances, actions, and contexts (presumably each day on Earth brings with it billions of events of relevant kinds, though each of us encounters only a few of them). And compared to application conditions, the notion of a C-role is a relatively blunt instrument. But it is I think essential to understanding quasi-disagreements and related issues.

8.1 C-Role versus Functional Role
We might wonder how C-roles for moral concepts relate to the output clauses of folk morality, and the functional roles they specify. In fact the application conditions/consequences distinction for concepts, though analogous to the distinction between those aspects of functional role specified by input and output clauses, is importantly different from it. Let’s call a functional role specified by Ramsifying folk morality an FM-role, and let’s call the part of that role specified by output clauses an FMO-role. Here are some observations.

(1) An FMO-role, like the complete FM-role of which it is a part, is a role for a descriptive property, whereas a C-role is a role for a concept.

We could in addition say: FMO-role-specifying output-clauses specify certain (noncausal)\textsuperscript{17} consequences of things instantiating the properties of goodness, rightness, and so on. An illustrative example from Jackson: properties that make something good are “the properties we typically have some sort of pro-attitude towards” (Jackson 1998a, 131; italics added). (This is an output clause because it licenses an inference from a claim explicitly about moral properties to a claim not explicitly about them.) As I read Jackson here, it
is the generality of instances of the relevant properties toward which we typically have some sort of pro-attitude, not some especially local subset of those instances.

C-roles, however, selectively relate individuals in possession of the relevant concepts, and their applications of those concepts, to particular cognitive, affective and behavioral events, in the lives of those very individuals: there is an inherent reflexivity in the notion. (In addition, pending a contrary ruling from some more substantive theory of concepts, I am taking the consequences in question to be causal consequences.)

(2) There’s a reason Jackson’s account could easily lead to confusion over these issues (and which makes what I said under point (1) above something of an oversimplification). If we were to Ramsify physical theory en route to clarifying the semantics and ontology of physics, we wouldn’t expect the output clauses dealing with ‘electron’ to state relationships between electron-beliefs and other phenomena: we would expect them to state relationships between electrons and other phenomena. Yet when we look at what Jackson says about “output clauses” for folk morality, virtually without exception, he mentions relationships between moral judgments and other phenomena, rather than relationships between moral properties and other phenomena. The example quoted in point (1) above is the only exception. But even that example comes from the following passage (which, notably, comes fairly early in the general presentation of moral functionalism in chapter 5): “The output clauses of folk morality take us from ethical judgments to facts about motivation and thus behavior: the judgment that an act is right is normally accompanied by at least some desire to perform the act in question; the realization that an act would be dishonest typically dissuades an agent from performing it; properties that make something good are the properties we typically have some kind of pro-attitude towards, and so on” (Jackson 1998a, 131).

Since these output clauses concern not ethical facts but ethical judgments, one could almost take Jackson here to be talking of folk psychology rather than folk morality! There has to be a prima facie worry here that Jackson’s version of moral functionalism has gone astray. The idea is that the moral properties are in principle identifiable as those that (as Jackson writes) ‘will/would’ play a certain complex set of interrelated (noncausal) roles. How do the (causal) roles played by ethical judgments help us to identify ethical properties? Judging that someone is dangerous will normally lead me to treat him with caution, whether or not he is actually dangerous: similarly, wouldn’t my judgment that something is right normally lead me to be inclined to do it, whether or not it is actually right?
In fact Jackson here is violating a constraint Lewis lays down in “How to Define Theoretical Terms”: “We must assume that all T-terms in the postulate of T are purely referential, open to existential generalization and to substitution by Leibniz’s Law” (Lewis 1970, 429). Fully to explore this issue might require an investigation of the notion of de re desire (or de re “pro-attitudes”) including such uneasy questions as whether and how one can have de re attitudes to properties as such.

This is not the place to explore such technicalities. What should be noted, though, is that these passages actually reflect Jackson’s commitment to a response-dependent account of moral facts, noted earlier. It is typical of response-dependent accounts to underwrite unusually tight connections between judgments and what they concern. But there is a complex array of considerations lurking here to do with the ontology and epistemology of “response-dependent properties.” For starters, the notion of response-dependent properties is much more problematic than the notion of response-dependent concepts. Arguably the relation of concepts to properties is a many–one relation; and arguably response-dependent properties are not so inherently, but merely qua how suitable individuals respond to them under suitable conditions. In any case, response-dependent concepts will track determinate properties only to the extent that the relevant conditions are determinate. There is a great variety in the kinds of conditions that may be relevant—some response-dependent concepts relate to responses under conditions conceived of as “normal,” some relate to responses conceived of as “ideal,” and either way, the nature and state of the respondents as well as of their environment at the time of responding will typically need to be included among the relevant “conditions.” Thus the epistemology of response-dependent properties is inextricably interwoven with the epistemology of relevant (in this case, “ideal”) conditions. We may desire to say that a proper specification of those conditions should be geared to the requirement that the relevant responses should “track” the properties we are interested in, but, as is well known, it is all too easy to fall into circularity in attempting to meet this demand. In the case of Jackson’s approach to moral functionalism, the “ideality” of the relevant conditions incorporates the notion of what would be desired or valued were one ideally rational, so the circularity that (as Jackson says) must be avoided is the one we would generate by defining ideal rationality as what we exhibit when our desires track genuine values.

Thus there is a complex web of issues to be untangled in evaluating the acceptability of Jackson’s inclusion among moral functionalism’s “output clauses,” of clauses relating moral judgments, rather than moral properties,
to motivation. Though we cannot deal with them all here, the issues, and the relevant passages, are important: for one thing, they reveal how deeply Jackson has woven the idea of a response-dependent account’s being endorsed by mature folk morality, into his whole account of moral functionalism. Once again, I am skeptical on at least two counts. One is skepticism about whether “mature folk morality” can be expected to find an agreed-on, noncircular, and determinate account of “ideal rationality.” Another, at least as central, is whether a reasonable account of the primary conceptual commitments of the folk can permit the content of some future or counterfactual best-case scenario for the outcome of critical reflection and debate—one in which the subtleties of response-dependence have all been successfully negotiated and incorporated—to be taken as trumping their salient, current, and firmly held (albeit philosophically less sophisticated) judgments about what is right or wrong, good or bad, in determining the content of those current judgments. What deepens the latter concern is the observation that just as on this account the folk’s actual moral judgments take second place, in delineating their actual moral concepts, to their merely counterfactual judgments, so too their actual moral motivations are similarly trumped by their merely counterfactual moral motivations.

That said, these elements in Jackson’s position do also amount to some kind of lessening of the prima facie gap between my view and Jackson’s. For instance, ‘the judgment that an act is right’ is not much different from ‘an application of the concept ‘right’’. But this should not be exaggerated. Jackson is saying that it is implicitly a priori in mature folk morality that there is (or will or would be) a connection of this kind between moral judgments and various pro-attitudes. This is to say that the existence of such connections is criterial for the applicability of moral concepts: it is not to say that possession of moral concepts is responsible for such connections. It is not apparent in what sense purely descriptive concepts could carry such responsibility.

(3) Jackson perhaps comes closest to what I say about C-roles—including the aspect just mentioned—in his discussion of the “Open Question” argument against ethical descriptivism, in particular where he moves from discussing the “content” strategy for explaining “the directed nature of the belief that A is right, within a purely descriptive framework,” to discussing what he calls the “content-possession” strategy for “adding motivation to direction.” Jackson says, inter alia: “on the moral functionalist story, to believe that something is right is to believe in part that it is what we would
in ideal circumstances desire. . . . Now this fact will typically manifest itself in our feeling to some degree the ‘tug’ of A” (Jackson 1998a, 159). (‘A’ is something believed right.)

Jackson continues in a postscript, considering a protest from “noncognitivists,” which is similar to (though stronger than) my own insistence on the importance of C-role. (I consider my own position, incidentally, a hybrid of cognitivist and noncognitivist elements, rather than straightforwardly cognitivist or noncognitivist). The protest “insists that when [I take this to mean ‘whenever’] one judges, really judges, that A is right, one must have a current, first-order pro-attitude towards A” (Jackson 1998a, 160), to which Jackson replies: “we can accommodate this view by refusing to call something a moral belief unless it is accompanied by the relevant pro-attitude” (ibid., 161).

This is one of the passages that lead me to say that “some may feel that little more than matters of emphasis separate us.” But there is still a significant gap. Note the difference between saying ‘a concept-tokening is not moral unless linked appropriately to motivation of its possessor’ (which is closer to what Jackson says) and ‘a moral concept is not possessed unless it is linked appropriately to motivation (etc.) of the possessor’ (which is what I say).20

8.2 “Harmony” for Ethical Concepts?
Returning to ethical quasi-relativism. How, it may be asked, do application conditions and C-role mutually constrain one another? Even in the case of a logical constant, what it takes for the analogues of these aspects (introduction rules and elimination rules) to dovetail suitably is debated—but at least the question lends itself to precise answers (see Read 2000). But that kind of “harmony” is very different from anything that could sensibly be proposed in ethical cases.

At this point I can do no better than quote Dummett himself. He is talking about personal identity, but what he says is mutatis mutandis pretty much applicable to ethics also, in my view. He says:

not to say that the character of the harmony demanded is always easy to explain . . . the most difficult case is probably the vexed problem of personal identity. An assertion of personal identity has consequences both for responsibility for past events and motives in regard to future ones. We can imagine people who employ different criteria for personal identity, but attach the same consequences to its ascription: what is difficult is to say where their mistake lies. (Dummett 1981, 455n)
And

If there were no connection between truth-grounds and consequences, then the disagreement between us would lie merely in a preference for different concepts, and there would be no right or wrong in the matter at all. Because we feel sure that there must be some connection, but we cannot give an account of it or say how tight it is, we are pulled in both directions: we feel that the disagreement is not a straightforward factual one, but that, at the same time, it is not merely a matter of applying different concepts. (Ibid., 358)

These are just the kinds of considerations that can lead one to a quasi-relativist view. If we could establish an objective, neutral, canonical, and precise criterion for an analogue of “harmony” between application conditions, and consequences (i.e., “C-role”), for moral concepts, then of course it would provide a basis for declaring at most one of those who agree in the C-roles for moral concepts, but disagree in the application conditions, correct, others being mistaken. But to ask what is the right “harmony” relation between application conditions and C-role, if I am right, is just a trivial variation on asking to know which moral judgments are correct. We should apply Dummett’s remark, that “we feel that the disagreement is not a straightforward factual one, but at the same time, it is not merely a matter of applying different concepts,” to the moral case: the disagreement, we should say, is a moral one, and such disagreements are not straightforward factual disagreements; but nor are they merely a matter of applying different concepts. We cannot sensibly say these things unless we view moral concepts as not entirely constituted or individuated by their application conditions, or their purely cognitive role, but as combining two aspects that cannot vary quite independently, though there can be a significant amount of “slippage” between them.

One classic feature of moral debates that makes this possible is a pervasive phenomenon in matters, especially, of judgment and evaluation. Even if there is a measure of agreement about many relevant parameters of, or considerations relevant to, evaluation, in such situations there is typically no nonarbitrary way of measuring the relative strength or importance of those considerations, and no nontendentious way of combining them so as to always arrive at an agreed overall conclusion. Saying the proper way to do things is to use “critical reflection,” or to value things “in accordance with rationality,” seems to me to provide no real way of progressing in such circumstances.21

If moral truth consists only in what judgments or evaluations would be converged on under ideal circumstances, and if which circumstances count as ideal is underdetermined, or irresolvably contested, so there is no single
definitive and canonical notion of ideal convergence, then the truth about corresponding moral disputes must itself be vague, indeterminate, or irresolvable. Quasi-disagreements are here to stay.

There are two motives one may have for thinking about idealized moral consensus. One is to secure the idea of non-agent- or community-relative moral objectivity or moral fact, in the context of something like a response-dependent account of what constitutes such facts, despite the familiar and pervasive actuality of moral disagreement. The other, less noted, but logically prior to the previous, is to secure a common subject matter for moral disagreement to be disagreement about. But there is something very odd for the latter purpose about substituting for people’s actual working moral concepts, an account of concepts they don’t but supposedly would all share in some perhaps impossible counterfactual circumstance. At best this would seem to leave us merely with failures to disagree combined with some in-principle commitment to negotiating common concepts in terms of which people might counterfactually express as genuine the disagreements they would then no longer have.22

Acknowledgments

This essay has benefited from helpful comments on earlier versions from a number of people including a conference audience at the Australasian Association of Philosophy (New Zealand Division) Conference at the University of Waikato, in 2004, and seminar audiences at Monash University and the Australian National University, in 2005. Among those to whom I am especially indebted are Tracy Bowell, David Braddon-Mitchell, Yuri Cath, Andy Egan, Frank Jackson, Jonathan McKeown-Green, Michael Smith, and Kenneth Taylor. Though intended to be relatively self-contained, this essay is also in some ways a companion piece to my “Failing to Agree or Failing to Disagree?—Personal Identity Quasi-Relativism” (Robinson 2004).

Notes

1. This label is not intended to connote any particular relationship to what Simon Blackburn calls “quasi-realism.”

2. All page references will be to Jackson 1998a unless otherwise indicated. For present purposes I am assuming that that work distils whatever is relevant in Jackson’s earlier discussions of moral functionalism, including Jackson 1992, and Jackson and Pettit 1995, 1996.
3. I am under no illusion that anything I say here could be construed as the “last word” on positions or issues here discussed. I have left unanswered many obvious questions about ethical quasi-relativism, which is why I call it a “sketch” of a position. Nor have I tried to compare this view with other views springing in part from similar concerns, and I have not discussed the content or relevance of Jackson’s subsequent work on related topics, such as Jackson 1999.

4. Given what follows, it is important to be clear here that a “no-fault disagreement” need not be one in which neither party can find grounds to object to the other (that they may do is a plank of quasi-relativism). It is a disagreement in which, symmetrically, no grounds exist on which either party can object to the other, apart from grounds that are, in the context, question-begging.

5. The favored technical term is ‘platitudes’. See, e.g., Lewis 1972. In his 1994, Lewis renounces the appeal to platitudes, describing folk theory (in particular, folk psychology) as “tacit” common knowledge, akin to grammatical knowledge.

6. Suppose I have a friend who insists that at least for her, it is analytic that someone is running only if in rapid bipedal motion in which their heels fail to touch the ground. I take her to a sports field and ask her to point out those who are running. Several of those pointed out are repeatedly making contact with the ground with their heels. My friend’s supposed concept of running has been shown to be “epiphenomenal” with respect to her practice.

7. The propriety of sometimes adjusting peripheral claims to better fit with core claims will be relatively straightforward in the first sort of case, provided a “realist” rather than a “reductionist” attitude is taken to the theoretical entities mentioned in the “core.” Things may be less clear in the second kind of case, since there a real “direction of fit” issue may arise regarding the relationship of supposed “core” to “periphery.” If relevant general principles are best seen simply as summarizing the more particular claims that fall under them—whether in the manner of exceptionless generalizations, or in the manner of ‘as a rule’, ‘for the most part’ generalities—then it will be inappropriate to modify the more particular claims so as to better fit the principles belonging to the putative “core.” That would be proper only to the extent that we had some degree of an independent epistemic handle on the generalizations as such. Otherwise the particular claims would be the “core.” Though I have little space to explore it, this is a highly contestable point in ethics, one that has substantial relevance to what follows.

8. Nothing in this section should be interpreted as expressing wholesale skepticism about the method of cases, which I hold to be well-nigh indispensable for many philosophical purposes. See Robinson 2004.

9. Consider folk belief in witches. A physicalistic philosopher might produce a “cleaned up” version of folk witch theory that omits the parts about casting spells and performing black magic, on the grounds that such activities are known to be
impossible. The resulting theory might be satisfied by various harmless if feisty women with an interest in herbalism. It would not be well, however, to declare those women to be witches. A good part of the reason for this is that it would be unreasonable to attribute a commitment to the cleaned up theory being a correct theory of witches, to paradigm folk users of the word ‘witch’.

10. See also note 7 above.

11. I don’t mean the example to be one where Bargle once played a version of the game with deviant rules. He is simply, let us say, easily confused, and has consistently misinterpreted a long series of questionable umpires’ rulings.

12. I have put these points in simple generic terms—since any good theory of reference should accommodate them. Readers may wonder how well they gibe with the “two-dimensionalist” semantic framework that Jackson has espoused and promoted. The answer I believe is “quite well,” and it is I think interesting, relevant, and illuminating to see how. Space prohibits saying more here, but what I have to say can be found in Robinson 2004, 526–528.

13. Among other implications, this means that an issue raised by note 7—what is the proper “direction of fit” between intuitively endorsed particular moral judgments, and intuitively endorsed general moral principles?—remains very much on the table given Jackson’s views. Major questions, not here able to be addressed, about the epistemology of various kinds of response-dependent notion, arise in this connection. But I have a hunch that the risk of intuitions being a product of “epiphenomenal” theorizing is if anything higher in the case of relatively concise general moral principles than in regard to particular case-by-case moral judgments.

14. In practice these two parts are mingled in a test-and-correct process, oscillating between refining the division of cases, and testing and improving hypotheses about principles governing the division.

15. Not to omit the possibility that a relevant “other tongue” might be or include some purpose-built product of science, philosophy, or mathematics, rather than being one of the world’s “natural languages” (think for example of “probability”).

16. Compare Lewis: “‘Qualia’ isn’t a term of ordinary language. Neither is ‘phenomenal character’ nor ‘raw feel’ nor ‘subjective quality’... Yet despite the lack of a folksy word or phrase, I still say that the concept of qualia is somehow built into folk psychology” (Lewis 1995, 140).

17. Jackson (1998a, 131) says moral functionalism’s principles “are not causal principles... The principles of folk morality tell us which properties typically go together, but not by virtue of causing each other.”

18. What Jackson most explicitly endorses is a response-dependent account of values, but it is clear that he sees moral facts as crucially involving suitable values;
I base what I say here on this fact together with his choice of sample “output clauses.”

19. This is very tricky. Jackson (1998a, 131) gives as an example “properties that make something good are the properties we typically have some kind of pro-attitude towards” (my emphasis). But ‘typically’ seems wrong here given the role of “mature” folk morality, in Jackson’s account. If current folk morality, by the lights of mature folk morality, currently gets it wrong about some properties, then these will be properties that do not after all make something morally good, though current folk morality may mistakenly think so. Furthermore, if current (or mature?) folk morality is to consistently invoke the notion of rationality in the way Jackson suggests, we would expect the clause to take a form more like ‘properties that make something good are the properties those who are ideally rational typically have some kind of pro-attitude towards’. But perhaps ‘good’ in this context does not mean only ‘morally good’.

20. This is a prima facie difference, premised on the view I am taking of concept individuation. Imagine once again the psychopath who can mimic the application of the predicates ‘right’ and ‘wrong’ but is devoid of moral motivation. I say he or she lacks moral concepts. But someone who possesses moral concepts may still in my view count as making moral judgments in applying them, even if on some of those occasions, moral motivation is absent. Jackson’s talk of “current, first-order pro-attitudes,” however, suggests that a person employing the same concept (descriptively individuated) on different occasions may sometimes make moral judgments in applying it, other times not, as his or her motivation waxes and wanes.

21. One of the most obvious—and difficult—questions I have left outstanding about ethical quasi-relativism is what explains the constraints on “slippage” between application conditions and C-roles for moral concepts. Something less superficial—and less local to a particular language—is needed than a mere fact about correct usage of words like ‘morality’. Perhaps the model of a set of agreed relevant considerations, but lacking quantifiability or uncontentious algorithms for combination, might form part of the story in accounting for the constraints. But that’s a mere conjecture.

22. Much of what I have to say about what I call “quasi-disagreements” has points of similarity with views discussed (in much more detail), and in some cases endorsed, in the excellent book Tersman 2006. Folke Tersman calls such disagreements “radical disagreements.” This essay was completed and submitted before Tersman’s book, from which this essay might otherwise have benefitted, came out.
References


Contributors

David Braddon-Mitchell
Department of Philosophy,
University of Sydney

Mark Colyvan
Department of Philosophy,
University of Sydney

Frank Jackson
Departments of Philosophy,
Princeton University and LaTrobe
University

Justine Kingsbury
Department of Philosophy,
University of Waikato

Fred Kroon
Department of Philosophy, The
University of Auckland

David Lewis
David Lewis was Class of 1943
University Professor of Philosophy
at Princeton University

Dustin Locke
Department of Philosophy,
University of Michigan

Jonathan McKeown-Green
Department of Philosophy, The
University of Auckland

Kelby Mason
Department of Philosophy, Rutgers
University

Peter Menzies
Department of Philosophy,
Macquarie University

Robert Nola
Department of Philosophy, The
University of Auckland

Daniel Nolan
Philosophy Department, University
of Nottingham

Philip Pettit
Departments of Philosophy and
Politics, Princeton University

Huw Price
Department of Philosophy,
University of Sydney

Denis Robinson
Department of Philosophy, The
University of Auckland

Stephen Stich
Department of Philosophy and
Center for Cognitive Science,
Rutgers University

Daniel Stoljar
Philosophy Program, RSSS,
Australian National University
Index

Abduction, 16, 232
Adams, Robert M., 221
A-intension, 37, 226, 240
Aliens, 205–206, 212–213, 216, 221
Aquinas, St. Thomas, 126
Armstrong, D. M., 99, 105, 111, 208, 220–221
Bayesian, 290, 292, 300, 303, 308–309, 312
Behaviorism, 1, 100, 134–135
Belnap, Nuel D., 216
Best deserver, 269
Biconditionals, a priori, 12, 99, 101, 103, 105, 107, 109, 111
Birds, animatronic, 36
Black, Robert, 209, 218, 221, 240
Block, Ned, 109, 161, 254
Boghossian, Paul, 198
Braddon-Mitchell, David, 11, 30, 40, 111, 143, 156, 238, 280, 345
Burge, T., 274
Byrne, Alex, 161
Caloric, 273–274
Capacity, recognitional, 107
Carnap, R., 3, 4, 19, 20, 186, 208, 280–281
Carnap sentence, 186, 191, 208, 220, 280, 281
Carruthers, P., 258
Causal powers, 203
Chalmers, David, 69, 74, 131, 133, 137, 141, 157–158, 161, 222, 244, 265
Child, William, 181
Chomsky, Noam, 53–54, 56, 66–67, 96, 175, 177
C-intension, 222, 226, 232, 240
Circles, recognition of, 60, 65, 68, 83, 89, 93
Cognitivism, 318–319
Collins, J., 19
Colyvan, M., 18, 306, 312–313
Combinatorialism, 16, 208–209, 212–213, 220–221
Compatibilism, 99, 101, 103, 105, 110, 379
Competence, semantic, 14, 175–180, 182, 273, 321
Composition, 106–107, 110, 117
Conceivability, 133–136, 251
theory of, 35, 38, 340
Conceptual analysis, 14, 23, 47, 94, 160–182, 199, 267–271, 283–287, 295–298, 335
Conceptual change, 40–41
Consequences, 31, 33–34, 162, 194, 267, 287
logical, 206
long-term, 34
short-term, 34–35
Conservatism, epistemic, 290–295
theoretical, 290, 295
Constitution, 102, 270
“is” of, 102
Contextualism (about knowledge), 234–236, 240, 241
Convergence, 19, 34–35, 324–325, 333, 345
Coordinates, 17, 243–256, 261
Core-physical, 104
Cummins, R., 253

Davidson, Donald, 162
Davies, Martin, 131, 156, 180, 240
Decision theory, 18, 254, 266, 303, 306, 308, 310, 312
Deduction, 293, 337
Dennett, D., 257

Derivation, 17, 244–246, 248–250, 252, 256, 266, 304
Descriptions (definite), 4–6, 10, 20, 139–158, 176, 191, 211, 216, 224, 230–231, 233–234, 237–238, 267, 284, 286, 331
Descriptivism, 14, 139–158, 280–283, 319, 332, 342
Determinism, 101, 279
Devitt, Michael, 144, 150–151, 176
Disagreement, 26, 40–41, 47, 316–345
moral, 18–19, 322–326, 328, 345
Dispositions, 11–12, 38–40, 42, 177, 179, 255–256, 262, 396
Dretske, Fred, 220
Dummett, Michael, 337–339, 343–344

Electron, 3–6, 103–104, 255, 257, 268, 276, 282, 299
Eliminativism, 26, 63, 111, 196
Ellis, Brian, 220
Equivalence meaning, 157
Essence, 24, 115–123, 126–134, 270
Eucalyptus, 36

Fine, Kit, 120, 221
Free will, 7, 37, 101–102, 105, 110, 113, 175, 265, 267
Frege, Gottlob, 154, 156–158

Genes, 25, 269, 287
Gettier, Edmund, 163, 165, 167, 235, 335
Gödel, K., 312
Goodman, N., 291–292
Gopnik, A., 96
Gore, Al, 226
Grammar, generative, 54–55, 66
Grice, Paul, 160, 163, 165, 170, 176
Haecceitism, 209–212, 221, 227, 240
Hájek, A., 312
Harmony, 337, 343–344
Haugeland, John, 181
Hawthorne, Jonathan, 1, 114, 299
Heat, 128–130, 133, 136, 222, 274, 276, 285
Hempel, Carl, 2, 19–20
Hintikka, J., 311
Hume, D., 263, 303, 309
Humanism about laws, 103, 269
Humility, Kantian, 203
Ramseyan, 16, 203–222, 223–225, 240
Hyperintensionality, 29, 37
Idealization, 34–35, 65, 75, 77, 333
Identity, transworld, 210–211
Idlers, 16, 205–206, 212–216, 235, 239–240
Input, 30–31, 34–35, 162, 255, 324, 339
Introspection, 14, 40, 52–53, 58–59, 62, 287
Johnston, Mark, 114, 116, 321, 333
Kahneman, D., 312
Kinds
functional, 41
natural, 104, 139, 172, 274, 281, 328
social, 329–330
Kind-terms, 141, 145, 157, 280
Knowing-that, 119, 133, 225, 232, 238–239, 248
Knowing-what, 48, 118–119, 126, 215, 238
Knowledge, tacit, 48, 118–119, 126, 215, 238
Kolmogorov, A., 303, 312
Kroon, Frederick, 13, 14, 141, 156–158
Lakatos, Imre, 175
Langton, Rae, 203–204, 214, 223, 225, 228, 235
Language, Lagadonian, 211–212
Laurence, S., 36
Laws (of nature), 6, 27, 103, 108, 162, 188, 203, 205, 208–209, 213, 218–221, 267, 269, 272
Levine, J., 226
Libertarianism, metaphysical, 33
Location problem(s), 9–10, 195, 267, 269–271, 279, 332
Locke, J., 286
Logic, classical, 303, 306, 308, 312
Logic, Kleene, 306
Lowe, E. J., 112
Lycan, William, 171–173

Mapmaking, 34
Meno, 65, 143
Mental states, 8, 14, 18, 30–31, 33, 35, 40, 46, 84, 159, 166, 171–172, 174, 187–180, 268, 276, 281, 283
Menzies, Peter, 14, 15, 164, 183, 191, 192
fundamental, 18, 100, 102–104, 110, 160, 176, 269, 283, 298
Method of cases, 180, 319–321, 346
Mill, John Stuart, 156, 220
Miller, K. L., 30
Millikan, Ruth, 175
Momentum, 192, 285–286
Morality, mature folk, 19, 191, 323–324, 326, 331–332, 338, 342, 348
Müller-Lyer illusion, 258
Multiple realization, 4

Nagel, T., 258
Names, 5–6, 13–14, 20, 139–158, 191, 206, 212, 214, 217, 230, 280, 329
Naturalism, 23–28, 174–175, 306–312
Nichols, Shaun, 84, 96
Nola, Robert, 302, 308
Nolan, Daniel, 7, 18–20, 26, 27, 280
Normativity, 11, 18, 32, 296, 303, 305–311

Occam’s razor, 205, 213, 239
O-language, 205–207, 217, 219, 224
Ontology, 9, 13, 18, 26–27, 47, 172, 174, 256, 270, 296, 340–341
moderate, 102
O-terms, 2–6, 11, 185–187, 194, 224, 230
Output clauses, 333, 339–341, 348

Papineau, David, 305
Personal level, 49, 50–53, 57–59, 62–64, 71
Pettit, Philip, 17, 182, 243, 244, 248, 256, 258, 262, 263, 264, 345
Phonemes, 67
Physics, quantum, 48
Plantinga, Alvin, 221
Possibilia, 23, 25–26
Price, Huw, 1, 14, 114, 136, 299
Prior, Elizabeth, 216
Probability, 18, 187, 254, 290, 303, 312, 347
Properties
alien, 16, 221
perfectly natural, 218
physical, 43, 159, 183
semantic, 40, 151, 195, 198
Prototype theory of concepts, 35
Putnam, Hilary, 167, 231, 274
Quasi-realism, 345
Quidditism, 16, 208–213, 221, 223
causal theory of, 141–142, 145–147
description theory of, 154
Referential indeterminacy (ambiguity), 20, 196–197, 220–221
Relativism, quasi, 19, 315–316, 325, 334, 343, 346, 348
Replacement, 210, 213, 135, 239
argument, 212, 214, 239
conception, 13, 113–114, 124–125, 136, 275
Representational states, 17–18, 31, 71, 252–261, 265
Research School of Social Sciences, 1, 25
Revelation, doctrine of, 113–136, 115, 124, 129, 135
Rigidify (rigidification), 43, 139, 157, 216, 222, 231, 233–234, 238
Robinson, Denis, 182, 220, 346, 347
Roles, nomological, 204, 207–208, 220
Rules, generative, 54–55, 67
Russell, Bertrand, 154, 156–158, 222, 312
Salmon, Nathan, 158
Schiffer, Stephen, 163
Searle, John, 141
Segerberg, Krister, 161
Self-presentation, 116–117, 122
Index

Sentences, grammatical, 5, 9, 60, 67–69, 73–74, 76–78, 80–81, 83, 89, 91, 93, 96
open, 60–61, 66, 69–70, 79, 88–89, 92, 224, 322
Shoemaker, Sydney, 220
Similarity between aggregations, 100, 106
Simulation, off-line, 30
Skepticism, 16–17, 115, 228–229, 232, 234, 236, 238, 241, 342, 346
Smart, J. J. C., 28, 99, 105, 111
Smith, Michael, 181, 262, 299
Soames, Scott, 24, 141–144, 147, 150–158, 240
Space, logical, 100–104
Stalnaker, Robert, 48, 52, 58, 102, 109, 131, 133, 161, 222
Stanley, Jason, 139
Sterelny, Kim, 144, 150–151, 173–175
Stich, Stephen, 11–12, 46, 84–85, 96, 166, 196, 198
Stoljar, Daniel, 13, 20, 49, 120, 135, 265
Strawson, G., 116
Subpersonal, 49–53, 57–59, 62–64, 71, 156, 257
Supervenience, 9, 106, 163, 276, 319, 333
Swoyer, Chris, 220

Tersman, Folke, 348
Theoretical terms, 4–5, 8, 15, 18, 23, 25–26, 139, 183, 185, 187, 192–193, 205–208, 220, 231, 273, 280, 282, 299, 308, 310, 312, 322
Theory, tacit or implicit, 12, 49–50, 53, 55–57, 61–64, 70, 72, 76, 84, 86–91, 94–95, 396
Thoughts, fine-grained, 30

Time-slice, 41–43
Tooley, Michael, 183, 191, 220
Triangle, 101–102
Truths, analytic, 39, 274, 278, 280, 282–283
Twin earth, 25, 39, 50, 58, 165
Two-step, Canberra Plan, 7, 9, 14, 27, 269–270

Unger, Peter, 222
Universals, 56, 100, 205, 210, 219

van Fraassen, B. C., 300, 312
von Frisch, K., 107, 109

Wason, P. C., 312
Water and H2O, 24, 36, 39–43, 50, 100, 107–108, 131–132, 142, 146, 165, 177, 188, 328, 331, 335
Wettstein, Howard, 158
wffs, 72, 75–76, 80, 83, 85–86, 89, 91, 95
Whittle, Ann, 223, 225, 227, 240
Williamson, Timothy, 181
Wittgenstein, L., 81, 270, 329
Wright, Crispin, 299

Yablo, Stephen, 161

Zebras, 32–33