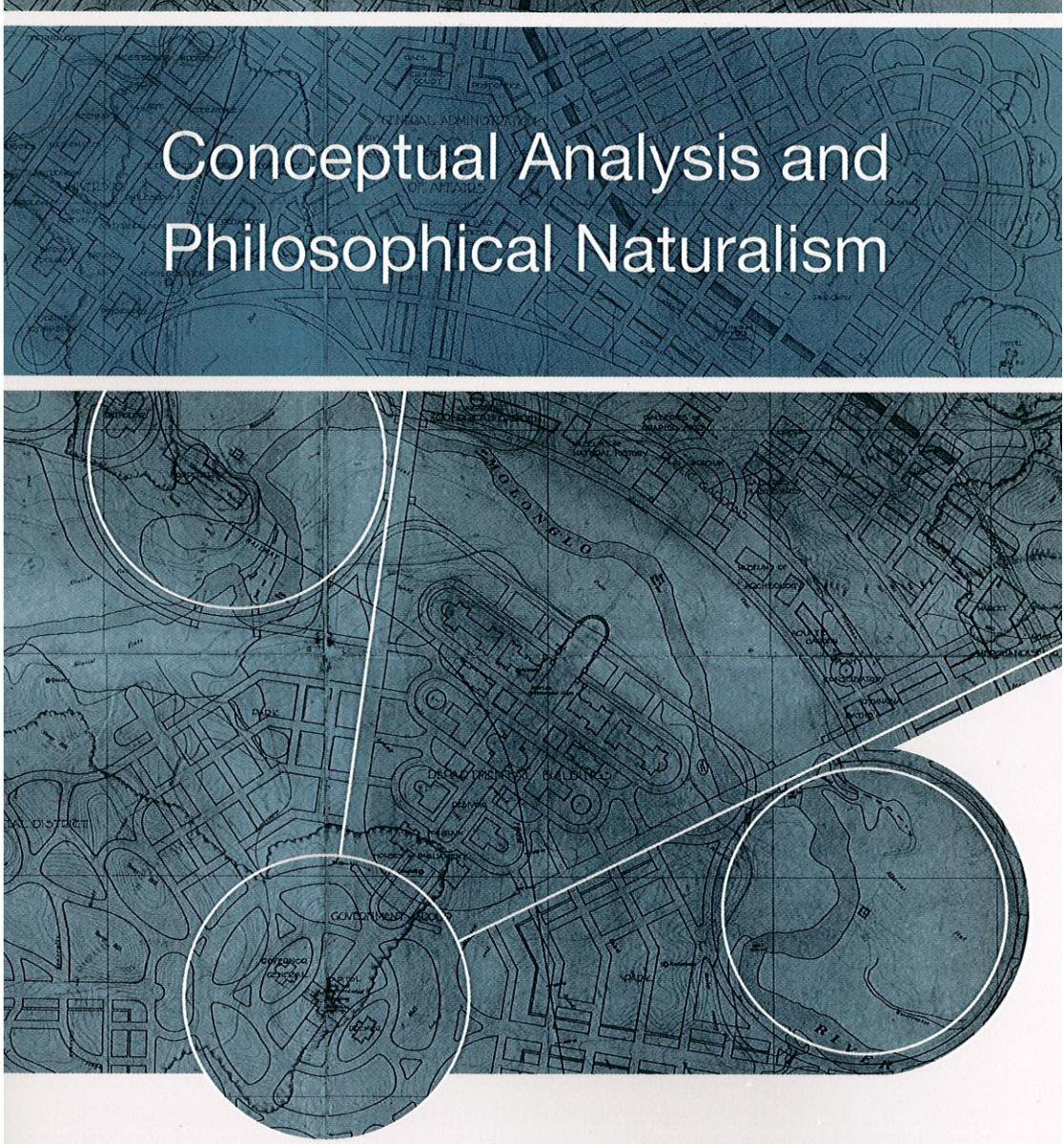
Conceptual Analysis and Philosophical Naturalism



Conceptual Analysis and Philosophical Naturalism

edited by David Braddon-Mitchell and Robert Nola

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

MIT Press books may be purchased at special quantity discounts for business or sales promotional use. For information, please email special_sales@mitpress.mit.edu or write to Special Sales Department, The MIT Press, 55 Hayward Street, Cambridge, MA 02142.

This book was set in Stone sans and Stone serif by SNP Best-set Typesetter Ltd., Hong Kong, and was printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Conceptual analysis and philosophical naturalism / edited by David Braddon-Mitchell and Robert Nola.
p. cm.
"A Bradford book."
Includes bibliographical references and index.
ISBN 978-0-262-01256-0 (hardcover : alk. paper)—ISBN 978-0-262-51228-2 (pbk. : alk. paper)
1. Analysis (Philosophy). 2. Naturalism. I. Braddon-Mitchell, David. II. Nola, Robert.
B808.5.C557 2009
146'.4—dc22

2008027190

10 9 8 7 6 5 4 3 2 1

Dedicated to the memory of David K. Lewis

Contents

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.[2] 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.[3] 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.[4] 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[5]). 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 modified 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."[8] 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"[9] 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 realizes 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 extra-scientific 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 Ramsification 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 1966, chapters 26–28.

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 of philosophical analysis are contained in Nolan 2005, 213–227.

10. For an account of what Lewis means by 'natural properties' see Lewis 1999, 13. 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