Synthese
https://doi.org/10.1007/s11229-020-02968-7

Troubles with the Canberra Plan

Panu Raatikainen[1]

Received: 9 October 2018 / Accepted: 19 November 2020

Abstract
A popular approach in philosophy, the so-called Canberra Plan, is critically scrutinized. Two aspects of this research program, the formal and the informal program, are distinguished. It is argued that the formal program runs up against certain serious technical problems. It is also argued that the informal program involves an unclear leap at its core. Consequently, it is argued that the whole program is much more problematic than its advocates recognize.

Keywords Canberra Plan · David Lewis · Conceptual analysis · Ramsey sentences

1 Introduction

The general approach commonly called "the Canberra Plan" is a rather influential research program in philosophy. At its core is a very specific view of the method of philosophy and philosophical analysis. The program is inspired by the systematic thought of David Lewis, and it gives an important role to the formal technique of "Ramsey sentences," also known as "the Ramsey–Carnap–Lewis method." The program's name derives from the fact that many of its central figures have had connections with the Philosophy Program of the Research School of Social Sciences at the Australian National University (ANU) in Canberra.[1]

[1] In fact, the label ("the Canberra Plan") was first introduced by critics (namely, O'Leary-Hawthorne and Price), and its original tone was not exactly flattering: "Canberra's detractors often charge that as a planned city, and a government town, it lacks the rich diversity of 'real' cities. Our thought was that in missing the functional diversity of ordinary linguistic usage, the Canberra Plan makes the same kind of mistake about language." (O'Leary-Hawthorne and Price 1996, p. 291, n 23) The advocates of the program have, however, adopted the expression and now use it to name their approach. Accordingly, it is no longer necessarily a negatively loaded name (cf. Nolan 2010).

✉ Panu Raatikainen
panu.raatikainen@tuni.fi
https://www.tuni.fi/en/panu-raatikainen;
https://philpeople.org/profiles/panu-raatikainen;
https://scholar.google.com/citations?user=p4gs2R4AAAAJ&hl=fi

[1] Faculty of Social Sciences, Pinni B4147, Tampere University, 33014 Tampere, Finland

Published online: 09 December 2020

Schematically, the program can be described as proceeding, in the case of any particular concept or family of concepts to be analyzed, in three steps.[2]

First, the theory essential for the concepts at hand must be identified. In the case of theoretical scientific concepts, one focuses on the scientific theory ("the canonical theory") in the context in which these concepts are first introduced ("defined"). In the case of common sense or philosophical concepts, the "platitudes"[3] concerning the concepts of interest are collected together; these are the relevant truths about the topic that most competent speakers (perhaps implicitly) believe. They constitute the "folk theory" of the area. The idea in either case is that the relevant theory "implicitly defines" the concepts at stake by defining their theoretical role.

Second, the theory—be it a scientific theory or a folk theory—is formalized. Furthermore, the vocabulary of the theory is somehow divided into *internal theoretical* terms (*T*-terms), introduced by the theory, and *observational*, *old*, or *outsider* terms (*O*-terms), which derive their meaning in some way external to the theory. The former are then "Ramsey-eliminated" or "Ramsified," and the Ramsey sentence and the Carnap sentence of the theory are achieved (more details below). The idea is that the Ramsified variant of the theory—that is, the Ramsey sentence of the theory—reveals the theoretical role of these concepts of interest.

Third, we look at the world (or our best current theory of it) in order to find out what in reality plays the role just described—that is, what realizes it. In this final phase, empirical science plays the main role. The earlier steps, in contrast, are done "in the armchair" by philosophers and, according to the advocates of the plan, result in *a priori* knowledge or conceptual truths.

In addition, at least Lewis (1994), and especially Jackson and Chalmers (Jackson 1994a, b, 1998; Chalmers 1996, 2012; Chalmers and Jackson 2001), take it that all this provides in particular an *a priori* entailment from microphysical truths to all ordinary macro-level truths (except perhaps phenomenal consciousness).[4] This, in turn, plays an important role in Chalmers' famous argument against materialism (e.g. Chalmers 1996).[5] Therefore, the philosophical stakes are quite high here.

1.1 Formal and informal Canberra Plan

Though in practice, the Canberra Plan has indeed been intimately tied to the formal Ramsey–Carnap–Lewis method, it is possible to isolate an informal philosophical idea behind the program that is not essentially dependent on the formal details of the Ramsey sentence approach. That is, it is useful to distinguish two forms, or aspects, of

[2] Nolan (2009) and Braddon-Mitchell and Nola (2009), for example, distinguish only two steps; they lump my first and second steps together into one step. I prefer to distinguish them as separate steps.

[3] What exactly counts as such a "platitude" is in fact a highly non-trivial philosophical question. However, I shall not dwell on that but pretend—for the sake of argument—that this can be satisfactorily decided.

[4] This further view may not count "officially" as an essential part of the Canberra Plan. However, it is in any case an important and influential view in contemporary philosophy, and is very closely related to the Plan and depends on it. Therefore, it is natural to discuss it in this context too (see Sect. 4).

[5] It is not clear to me whether Chalmers should be counted without reserve as a Canberra Planner. Nevertheless, Chalmers' scrutability framework is, in his own words (2012, p. 362), "at least a close relative" of the Canberra Plan.

the Canberra Plan: (i) the formal program that makes essential use of Ramsey sentences and related formal tools; and (ii) the informal philosophical program that focuses on platitudes and folk theories, and the causal-functional roles implicit in them, and aims to discover more precisely what plays those roles (Nolan (2015) refers to apparently (more or less) the same thing as "generalized functionalism"). It is easy to see that the informal idea emerged as a generalization of analytic functionalism in the philosophy of mind, due to Lewis and others.[6] Both aspects of the program have been popular in recent metaphysics. Nevertheless, it is my aim in this paper is to argue that they both face severe problems.

To be sure, Canberra Planners are a heterogeneous cluster of philosophers, and beyond the very general, basic ideas of the Plan, it is often difficult to determine who exactly is committed to this-or-that more particular view. Nevertheless, I shall also discuss some more specific views closely related to the Canberra Plan that are held at least by some Planners or go naturally together with the Plan. However, it is emphatically not contended that all Planners—or even the majority—commit themselves to all of these views. Still, these views are philosophically interesting and influential. They are therefore worth discussing, and it is natural to examine them in this context.

The primary aim of this paper is, however, to evaluate critically the Canberra Plan in general, as summarized above in three steps, and not to provide a comprehensive critical exegesis of the various more specific views of Lewis or those of some leading Planners. Consequently, I shall largely set aside many other aspects, particular details and later developments of Lewis' views that do not appear to essentially belong to the Plan (although some are commented on briefly). Nevertheless, just when things begin to get interesting, Planners tend to be very brief and to defer to Lewis' classic papers (Lewis 1970, 1972). Therefore, a critical reader has often no choice but to look at those early papers for possible clarification and further detail. It is plausible to assume that unless stated otherwise, Planners follow Lewis in these matters.

The paper is structured as follows. In Sect. 2, the formal Canberra Plan and the Ramsey sentence method are reviewed; in Sect. 3, certain serious problems regarding it are discussed; in Sect. 4, the informal Canberra Plan is critically discussed; and in Sect. 5, I draw some conclusions.

2 Theories and their Ramsey sentences

Let us first run over the main lines of "the Ramsey–Carnap–Lewis method." The approach has its origin in Frank Ramsey's classic article "Theories" (1929/1931).[7]

[6] Menzies and Price (2009) further distinguish two versions of the informal program: a narrower version that restricts its attention to causal roles, and a more global one that deals with all sorts of "functional" roles, of which many are clearly not causal roles. They then scrutinize the status of semantic notions in this more comprehensive approach and raise doubts about the viability of the generalized program. I am sympathetic to many of their observations, but in this paper, I shall mostly focus only on causal roles—which I take to be the paradigms of the Canberra Plan—also for the wider program. It seems to me that our critical reflections, even though they proceed in different directions, are complementary and do not conflict. My aim is to challenge the narrower program; if that challenge succeeds, it certainly undermines the generalized program too.

[7] The paper was published only posthumously in 1931, but it was written in 1929.

Fundamentally the same idea was re-discovered, apparently first without any awareness of Ramsey's work, by Carnap in the mid-1950s (see Psillos 1999, 2000), and the approach became popular in the philosophy of science through a series of publications by Carnap (1958, 1959, 1963, 1966, 1975; see also Bohnert 1967; Maxwell 1970). In contemporary metaphysics, though, it was primarily Lewis' articles (1970, 1972) that have made it current.

To begin with, the Ramsey sentence approach presupposes the division of the (non-logical) language of the particular theory S at stake into two mutually exclusive classes: Ramsey himself talked only abstractly about "the primary system" and "the secondary system." Carnap, however, related this framework explicitly to the orthodox *observational-theoretical* distinction in the philosophy of science, and this has been since then the standard interpretation. Finally, although Lewis also aimed to define *theoretical terms*, he was more critical toward the traditional observational-theoretical dichotomy, and preferred to call the expressions in the former class just "old terms" or "original terms"—or simply "O-terms." The latter are, in Lewis' understanding, terms which are already understood, whereas theoretical terms—T-terms—are the new terms introduced with the theory in question. An "O-sentence" is a sentence that does not contain any T-terms, and any sentence that contains T-terms (it may also contain O-terms) is a "T-sentence." In the beginning (in Lewis 1970), Lewis' focus was on theoretical terms in science, but the later Lewis and Canberra Planners often interpret "theoretical" very widely to include all sorts of concepts occurring in philosophy and folk theories. Nevertheless, the framework of the philosophy of science has served as a model here.

The standard Ramsey sentence approach[8] focuses on theoretical *predicates* and related *second-order variables*. Lewis, by contrast, actually considered explicitly only *individual constants*, or *singular terms*, and *first-order variables*. However, it is important to recognize that the standard approach does *not* thereby assume the full-blown second-order logic; in reality only a two-sorted first-order language is used. In fact, Lewis himself contends that we can focus on singular names *because* some names can be assumed to denote *properties, relations* or *classes*; and that some amount of *set theory* is in any case necessarily required (Lewis 1970, p. 429). And if so, it is no more problematic to include predicates and predicate variables interpreted extensionally as denoting sets of individuals and sets of ordered-pairs (and n-tuples) of individuals. The more standard setting of Ramsey sentences with "second-order" variables and quantifiers is not a single bit more metaphysically committed, but simply makes things more transparent. There is a simple translation between the standard two-sorted and Lewis' one-sorted framework, and it is more a matter of convenience which one is used. Moreover, once we have theoretical predicates, singular terms can be subsumed under them: simply define $T_{\mathrm{t}}(x) \leftrightarrow_{\mathrm{df}} (x = \mathbf{t})$. Consequently, in what follows, the focus is on the standard general approach involving theoretical predicates.

Now the central idea of the Ramsey–Carnap–Lewis method is the following: Assume that the theory S is presented in a standard form with theoretical T-predicates $T_1, T_2, \ldots, T_{\mathrm{n}}$, and observational/old O-predicates $O_1, O_2, \ldots, O_{\mathrm{n}}$. The Ramsey sen-

[8] That is, the approach of Ramsey himself as well as of Carnap, and of the larger part of the literature focusing on Ramsey sentences referred to in what follows.

tence S^{R} of S is obtained by first replacing all the theoretical predicates with distinct second-order variables, and then, to the result of this replacement, prefixing the existential quantifiers with respect to those second order variables. Thus, if the original theory S is written as

$$S(T_1, T_2, \ldots, T_n, O_1, O_2, \ldots, O_n),$$

then the Ramsey sentence S^{R} of S is:

$$(\exists X_1)(\exists X_2)\cdots(\exists X_n)S(X_1, X_2, \ldots, X_n, O_1, O_2, \ldots, O_n).$$

After Ramsey's initial suggestion, others have demonstrated various nice logical properties of Ramsey sentences. Here are some important ones (cf. Psillos 2006):

S^{R} is a logical consequence of S.
S^{R} and S have exactly the same O-sentences as their logical consequences.
S_1 and S_2 have incompatible O-consequences if and only if S_1^{R} and S_2^{R} are incompatible.
S_1 and S_2 may make incompatible *theoretical* assertions, yet S_1^{R} and S_2^{R} can be compatible.
If S_1^{R} and S_2^{R} are compatible with the same O-truths, then they are compatible with each other.

Consequently, Ramsey sentences may seem to well suit the purposes of capturing the factual, or synthetic, contents of theories. Lewis (1972, p. 254) notes: "The Ramsey sentence has exactly the same O-content as the postulate [theory] of T; any sentence free of T-terms follows logically from one if and only if it follows from the other."

Lewis contends that T-terms can be defined with the help of O-terms and Ramsey sentences involving only the latter, and are thus eliminable. Nevertheless, for him, this does not undermine realism: "I am also *not* planning to '*dispense* with theoretical *entities*.' Quite the opposite. The defining of theoretical terms serves the cause of scientific realism." (Lewis 1970; p. 428; my emphasis) Thus, Lewis does not advocate a radical empiricism in which all there is to the truth of a theory is its empirical adequacy. Presumably, the same is the case with Canberra Planners.

Even if the philosophical outlook of Carnap and Lewis may have been quite different—Carnap apparently advocates instrumentalism[9] whereas Lewis is a scientific realist—they seem to share the conviction that the Ramsey sentence technique can be used to split the theory tidily into two parts: the analytic part that defines the meanings of the theoretical terms, and the synthetic or factual part in which these theoretical terms do not at all occur (see Lewis 1970, p. 427). The Ramsey sentence of the theory is taken to faithfully capture the latter and exhaust the factual content of the theory (more of the analytic part below).

[9] Some (e.g. Psillos 1999) rather interpret the later Carnap as a structural realist—but it makes no difference for our arguments here.

2.1 Lewis and *O*-terms

Lewis is justly critical of the absolute distinction between the observable and the theoretical as advocated by logical empiricists.[10] The distinction is certainly vague and at best relative; there is arguably no pure observation, as observation is always theory-laden to some degree. However, the problems with the distinction should not be exaggerated either.

As long as what is "observable" is relative only to *other* theories and does not depend on the theory S at stake, its relativity and theory-ladenness is not seriously problematic. Even if we grant the relativity and blurriness of the distinction, we are still inclined to talk in a particular context about the support, confirmation, and disconfirmation that a theory receives from *observation*—about the theory's empirical adequacy—that it does not entail any consequences falsified by observation, and so on. Some sort of notion of observability, even if somewhat relative and vague, is indispensable. We should not throw the baby out with the bathwater.

According to Lewis, a theoretical term is a term that is "introduced by a given theory T at a given stage in the history of science" (Lewis 1970, p. 428). O-terms are, in contrast, those terms already understood at that stage of history—we are already acquainted with their meanings. But we may further ask: How? Are they theoretical terms of an earlier theory T^* and defined by such a theory at some earlier stage of history? Are there some terms that are not so defined, however far back we go, but which are understood and introduced by essentially different means, perhaps in relation to observation and ostension? If not, we end up with global descriptivism or global structuralism, and the language threatens to lose any connection to non-linguistic reality.[11] If so, we can say that those other terms are, roughly, observational ones.

In any case, however we draw the distinction,[12] we can call a theory *O-adequate* if it does not entail any false O-sentences. In general, there may exist several mutually incompatible theories that are equally O-adequate—perhaps they even have exactly the same O-consequences. We need some such notions if the many nice properties of Ramsey sentences listed above are to have any use.

2.2 Carnap sentences

Carnap made a further suggestion not to be found in Ramsey: he also paid special attention to the implications of the form $(S^{\mathrm{R}} \rightarrow S)$, where S is a theory and S^{R} is

[10] Lewis refers approvingly to Putnam (1962).

[11] Lewis himself warns about the disastrous consequences of such global descriptivism in another context (Lewis 1984). The view also faces the Newman objection (see below) in a particularly powerful form: then the Ramsey sentence of the ultimate theory is almost trivially true, if only there are sufficiently many individuals in the reality. Lewis (though he never explicitly discussed the Newman objection) makes essentially the same point in relation to Putnam's paradox: "[Global descriptivism] leads straight to Putnam's incredible thesis. For *any* world (almost), whatever it is like, can satisfy *any* theory (almost), whatever it says." (Lewis 1984, p. 60).

[12] It must be drawn somehow if the Ramsey sentence method is supposed to be used at all.

its Ramsey sentence; such an implication is nowadays standardly called the "Carnap sentence" of the theory S.[13]

The idea, originally suggested by Carnap, is that the Carnap sentence of a theory captures the *analytical component* of the theory—more exactly, that the analytical truths of a theory are exactly the consequences of its Carnap sentence. (Recall the dual idea that the Ramsey sentence in turn captures the synthetic or "factual" content of the theory.) Lewis apparently agrees with Carnap here.[14] He too notes the following important properties of Ramsey and Carnap sentences:

> (1) The postulate [theory] is logically equivalent to the conjunction of the Ramsey sentence and the Carnap sentence. (2) The Ramsey sentence and the postulate [theory] logically imply exactly the same O-sentences. (3) The Carnap sentence logically implies no O-sentences except logical truths. (Lewis 1970, p. 431)

2.3 Lewis' modified Ramsey sentences

Lewis in fact further defined what he called *modified Ramsey sentences*. These are just like Ramsey sentences except their quantifiers involve the uniqueness condition: a modified Ramsey sentence also says that the formula $S(X_1, X_2 \ldots, X_n, O_1, O_2 \ldots, O_n)$ has a unique realization. (Lewis, again, focused only on quantifiers over individuals; but for the general case, we must consider second-order quantifiers.) That is, informally, instead of having just "There is an X such that $S(X, O_{1, \ldots})$," we now have "There is exactly one X such that $S(X, O_{1, \ldots})$"; it is commonplace to abbreviate this formally as[15]

$$(\exists!\, X_1)(\exists!\, X_2) \cdots (\exists!\, X_n) S(X_1, X_2, \ldots, X_n, O_1, O_2, \ldots, O_n).$$

Accordingly, Lewis also reflects on modified Carnap sentences formulated in terms of such modified Ramsey sentences. The early Lewis (1970) contended that if the uniqueness condition is violated, the related theoretical terms of the theory fail to refer to anything.

2.4 Later developments in Lewis

Lewis' views evolved and changed over time. On the one hand, he later supplemented his framework with the condition that, in contradistinction to plentiful gerrymandered classes, only *natural* properties are eligible to serve as referents (Lewis 1984, p. 65). On the other hand, Lewis later relaxed the uniqueness requirement. He has rather suggested that in the case of more than one realizer, the relevant term or concept has

[13] Lewis was apparently one of the first to use this label.

[14] In (Lewis 1997, p. 334), for example, Lewis says that we can factor a theory into two parts: an existential claim and a semantic stipulation; he then refers to (Lewis 1970). He obviously means the Ramsey sentence and the Carnap sentence of the theory here.

[15] The unabbreviated form, with one existential quantifier, of course, reads
$(\exists X)[S(X, O_1, \ldots) \,\&\, (\forall Y)(S(Y, O_1, \ldots) \rightarrow Y = X)]$
(and similarly for more than one existential quantifier).

an *indeterminate reference* (Lewis 1984, p. 59) or is *ambiguous* (Lewis 1994, p. 301). On both occasions, Lewis refers to Field's notion of *partial reference* (see Field 1973).

Some Planners may be willing to follow the later Lewis here. But in the works of many Planners, there is no trace of these later views, only references to Lewis' early classic papers (Lewis 1970, 1972). Some seem to want to stick to Lewis' uniqueness requirement, but others only follow the general line of the Ramsey sentence approach. For these reasons, I have not included the uniqueness condition as an essential part of the Canberra Plan, nor have I included Lewis' later views involving natural properties and ambiguity or indeterminacy. I shall briefly comment on each of these separately, but I am not treating them as essential parts of the Plan.

3 Problems with the formal plan

Canberra Planners seem to assume that the Ramsey–Carnap–Lewis method is a more or less uncontroversial formal tool that is not in need any kind of further elaboration or defense. The literature is brimful of vague gestures towards this technique, but as far as Canberra Planners go, it is impossible to find a more detailed discussion of it. Lewis' original papers, themselves far from comprehensive from today's perspective, remain the most thorough discussions of the method in this tradition.

Although Canberra Planners virtually never mention this, there are in fact a couple of serious problems with the Ramsey sentence approach (and, consequently, with the whole formal Canberra Plan) in the literature: first, the so-called Newman objection should be already quite familiar; second, there is the somewhat less well-known Scheffler objection. Canberra Planners have nevertheless remained strikingly indifferent to these problems. I will now briefly review these objections, and develop the arguments a little further.

3.1 The Newman objection

To begin with, there is a now quite noted objection to the Ramsey sentence approach, the idea of which goes back to Newman's (1928)[16] critique of Russell's version of structuralism—hence its name, the "Newman objection." It was given a new lease of life in contemporary debate by Demopoulos and Friedman (1985).

However, we need to distinguish two uses, or two versions, of the objection.[17] First is *the traditional Newman problem*, which was presented against the pure structuralism of Russell (Newman 1928), and also applies against a similar view entertained at some point by the early Carnap, as well as against the extreme version of the Canberra Plan with global Ramsification (i.e., in which *all* non-logical terms are Ramsified). The argument now is that the relevant specifications of the world, and the Ramsey sentences

[16] Obviously, Newman himself did not address Ramsey sentences, as his paper predates Ramsey's seminal paper (Ramsey 1929/1931). Nevertheless, the contemporary objection to the Ramsification in question owes so much to Newman's argument that it is appropriate to call it "the Newman objection"; cf. Demopoulos & Friedman 1985.

[17] It seems to me that Chalmers (2012), for example, though he mentions the Newman objection few times, fails to sufficiently recognize the difference between these two (see Raatikainen 2014).

in particular, are nearly vacuously satisfied, if only the world has a sufficient number of objects. This is a grave problem for global structuralism.[18]

Second, there is the Newman objection as formulated in the more recent literature, *the contemporary Newman problem* (as one might call it). It was initiated by Demopoulos and Friedman (1985), and analyzed in detail by Ketland (2004, 2009; see also Ainsworth 2009; Button and Walsh 2018, Ch. 3); the setting here is somewhat more sophisticated, and the objection is not presented against global structuralism but against structural realism and related views[19]: in this formulation, not all predicates are Ramsified away, only the theoretical ones. It is this version of the problem that is relevant to the Canberra Plan too.

The contemporary Newman problem is the following: Assume that a (consistent) theory S is just observationally adequate, or O-adequate (see above); note that some of its T-sentences could be blatantly false. Then the corresponding Ramsey sentence S^{R} is almost trivially true, provided that the domain only has a sufficient number of objects. Therefore, the objection continues, structural realism, which leans on the Ramsey sentence approach, collapses virtually into radical empiricism and cannot supply a genuine middle ground between standard scientific realism and empiricist antirealism (which is its announced aim). More generally, the proposed idea that the factual or synthetic content of a scientific theory is captured by its Ramsey sentence looks implausible: the Ramsey sentence S^{R} is true in numerous worlds in which the original unramsified theory S is false. Therefore, the contemporary Newman problem is a problem also for the formal Canberra Plan.

The general trouble is that—in contrast to radical empiricism—any even modestly realistic view about theories holds that there is more to truth than just O-adequacy: the theoretical part of a theory must add something substantial to O-truths. And many theories apparently do that. A Ramsey sentence, however, does not add anything to the O-truths except perhaps in some cases a constraint on the cardinality of the domain. There is therefore a genuine loss of factual content when one moves from a theory to its Ramsey sentence.

But how about "the modified Ramsey sentences" á la early Lewis? Perhaps they are not vulnerable to the Newman objection? The modified Ramsey sentences indeed behave differently—but they do not really succeed in circumventing the problem.

To begin with, as van Fraassen (1997) has noted, certain basic results of model theory cause troubles: if the theory only has an infinite model, the upward Löwenheim–Skolem theorem entails that it has models of every infinite size—models that have different a structure. Therefore, the modified Ramsey sentence is in all such cases simply false. Note that this does not require that the theory has *only* infinite models: it is another basic fact (a well-known consequence of the compactness theorem) that if a theory has arbitrarily large finite models, it also has an infinite model.[20]

[18] As we have already noted (see footnote 11), Lewis (1984) recognizes more or less the same problem. On the other hand, he seems to have been ignorant of (what I call) the contemporary Newman objection.

[19] More exactly, against the popular version of structural realism that explicates the notion of the "structural content" of a scientific theory with the help of the Ramsey sentence of the theory.

[20] In other words, the objection applies to all theories that do not determine an explicit finite limit for the size of the universe.

Yet what if we rule out stipulatively all abstract set-theoretical entities and uncountable infinite models, and focus exclusively on concrete, real world entities? There are still problems. Namely, essentially the same model-theoretic construction on which the contemporary Newman objection is based entails that the modified Ramsey sentence is false in all sufficiently large models,[21] even if the initial unramsified theory is true in some of these models. Again, it seems implausible that the modified Ramsey sentence in general successfully captures the factual content of the theory. In sum, the modified Ramsey sentences do not make the troubles go away.

3.2 The Scheffler objection

Moreover, as was noted above, there is another weighty problem for the Ramsey—Carnap–Lewis method, an objection we may call "the Scheffler objection." The idea was briefly suggested by Hempel (1958), developed in much more detail by Scheffler (1963, 1968), and culminated in Niiniluoto's conclusive though unfortunately little known defense of this argumentation strategy (Niiniluoto 1972, 1973; see also Tuomela 1973, 1974). This critical argument was revitalized in Raatikainen (2012).

The objection is, briefly, as follows: It can be shown that Ramsification can sometimes ruin *inductive* relations between theories and observation. This is because there are cases in which the Ramsey sentence S^{R} is a logical truth, although the original unramsified theory S is not. We may thus have, for example, a case in which observations confirm a theory and disconfirm a competing theory, but not the respective Ramsey sentences that are both logically true. Obviously, no observation can disconfirm a logical truth. Consequently, a Ramsey sentence S^{R} can be true in a world in which the original unramsified theory S is false and empirical evidence even disconfirms the latter. The assumption that the Ramsey sentence of a theory faithfully captures the factual or synthetic content of the theory is, also for this reason, not at all plausible (see Raatikainen 2012).

3.3 Corollaries for Carnap sentences

Recall that the proposal was further that the *analytic truths* of a theory are exactly the logical consequences of the Carnap sentence of the theory. Trivially, Carnap sentences themselves should then be analytically true.

In Raatikainen (2011),[22] however, a new corollary of the Scheffler objection for Carnap sentences was noted. Let S and S^{R} be as in the above (in the Scheffler objection). Consider then the Carnap sentence of S—that is, the implication $(S^{\mathrm{R}} \rightarrow S)$. As

[21] Let us illustrate this with a simple example with one quantifier: $(\exists!X)\,[S(X, O_1, \ldots)$. This modified Ramsey sentence is false in any model in which there is more than one set that satisfies the formula $S(X, O_1, \ldots)$ (with X free). But the model-theoretic construction of the contemporary Newman problem guarantees that every O-correct model with a sufficiently large domain is such: the non-O individuals can be quite freely permuted and substituted with each other while the structure is retained. This consequence has apparently not been noted in the literature on the Newman problem. (For the construction, see e.g. Button and Walsh 2018, Chapter 3).

[22] Raatikainen (2011) is a sequel to Raatikainen (2012), even if they appeared in print in the reverse order.

.

its antecedent is logically true, the whole Carnap sentence is now logically equivalent to the original (unramsified) theory S (it is true if and only if the latter is true). Consequently, a possible empirical disconfirmation of S—and there could be one—counts simultaneously as a disconfirmation of the Carnap sentence of S. However, Carnap sentences were supposed to be analytically true according to the view at hand here. Surely, however, it should not be possible for analytical truths to be empirically disconfirmed (assuming their meaning has not changed). Hence, the argument suggests that the popular view that Carnap sentences capture the analytical contents of theories must be wrong.

Let us also note a consequence of the construction leading to the contemporary Newman objection that has apparently not been previously noticed in the literature[23]: For every O-adequate theory S, there are a number of models in which S^{R} is true but S is false, and consequently, in which the whole implication ($S^{\mathrm{R}} \rightarrow S$)—i.e. the Carnap sentence of S—is false. Yet the latter was supposed to be analytically true, and certainly such truths should not fail in any model. Consequently, the proposal that Carnap sentences capture the analytical contents of theories again seems implausible.

3.4 Lewis' later views for help?

We have noted certain changes in Lewis' thought. One might hope that Lewis' later ideas would help circumvent the problems discussed above. First, could confining oneself to *natural* properties enable unintended interpretations to be ruled out, and thus save the Ramsey sentence approach? This is an intuitively appealing idea and may at first sound promising. However, it is difficult not to think of the notion of *naturalness* as a *theoretical* concept of our overall theory. Therefore, it would itself get eventually Ramsified away. But then, the Newman problem kicks in again, and the suggested condition imposes no real constraint at all.[24]

Second, could Lewis' suggestion that theoretical predicates are sometimes *indeterminate*, along the lines of Field's notion of *partial reference*, dissolve the problems? It does not seem so. Field's notion applies primarily to relatively closely related, partially overlapping concepts (or terms referring to such). Thus, Lewis himself writes:

> Note well that this is moderate indeterminacy, in which the rival interpretations have much in common; it is not the radical indeterminacy that leads to Putnam's paradox. I take it that the existence of moderate indeterminacy is not to be denied. (Lewis 1984, p. 223)

[23] So what is the difference between this new corollary and the problem discussed in the preceding paragraph? Like the Scheffler objection on which it is based, the latter focuses on specific worst-case scenarios, where the Ramsey sentence turns out to be logically true (that is not in general the case). The present problem, by contrast, holds for *all* O-adequate theories which allow sufficiently large domains, and is in that way more general.

[24] As a referee pointed out, some might disagree and contend that *naturalness* should be taken as a primitive notion and left unramsified. Suffice it to say here that I am quite skeptical of us having so clear and distinct an idea of *naturalness* that we could just take it as a primitive. It seems to me a quite theory-laden concept. Moreover, this would be an odd deviation from the general view of concepts essential for the Canberra Plan.

The indeterminacy resulting from the Newman problem is, by contrast, typically radical and voluminous: a theoretical predicate which is in all reason unequivocal would now refer to numerous completely distinct and unrelated sets or relations. And saying that almost all theoretical predicates are so extremely ambiguous seems just implausible or *ad hoc*. This is a systematic problem with the formal framework of Ramsey sentences, not with the equivocality of theoretical predicates in general. Lewis himself also took as an evident "Moorean fact" that our language has a fairly determinate interpretation (Lewis 1983, p. 47). One may wonder whether the radical and extensive indeterminacy that would result from the ambiguity move, if it were used as a general response to the Newman problem, is consistent with this.

4 The informal Canberra plan

4.1 Background: Canberra planners and multiple realizability

Many basic ideas of the Canberra Plan emerged in the heyday of the type-identity theory in the philosophy of mind, and Lewis was at the same time one of the leading advocates of the latter (see Lewis 1966). Since then, the popularity of the type-identity theory has somewhat waned. The key reason is the argument from *multiple realizability* from Putnam (1967). The idea that many mental and other higher-level properties are multiply realizable and therefore not type-identical with physical properties is quite popular in philosophy nowadays.

The theme of multiple realizability was already briefly touched upon above in abstract form in relation to Lewis' modified Ramsey sentences. Lewis himself, though, advocated the type-identity theory and never accepted that the argument from multiple realizability would undermine it.

Be that as it may, if we now move from Lewis to the Canberra Plan, there does not seem to be any official or shared view about multiple realizability and its implications. Possibly some advocates of the program commit themselves from the start to a strong and all-encompassing reductive view according to which any legitimate (apparently) higher-level property must be reducible to and be type-identical with a lower-level property and, ultimately, with a property at the fundamental physical level. From such a point of view, something like the informal Canberra Plan may appear as a promising further development, but then the program at least cannot be appealed to in support of the type-identity theory without circularity.

However, it seems that many sympathizers of the program are also more or less tolerant towards multiple realizability and the non-reductive view, or at least prefer not to commit their views to the denial of them. Furthermore, as I shall argue in short, concepts which are defined in terms of causal roles, so central for the Plan, tend to be by their very nature multiply realizable. In the rest of this article, I will reflect upon, among other things, how the informal Canberra Plan fares in the presence of multiple realizability.[25]

[25] To be sure, the argument from multiple realizability is not uncontroversial. Nevertheless, this is not the place to be absorbed in that issue. Many Canberra Planners grant multiple realizability at least to some degree.

4.2 The prospects of the informal plan

Perhaps it was simply a mistake in the first place to give the formal Ramsey sentence approach such a central role. Maybe the informal philosophical idea behind the Canberra Plan is nonetheless plausible. Never mind Ramsey sentences and such—could we just collect together platitudes, analyze interesting concepts in terms of their *causal roles* in our folk theories, and look for what in the physical world best fits the role?

An old and worn example of analytic functionalism concerns *pain*. The idea is that it could perhaps be "defined" or analyzed as X such that:

> X is caused by tissue damage, and X causes wincing, moaning, and avoidance behavior.

Whatever in the physical reality then plays this causal role *is* pain. Assuming this were an adequate analysis,[26] this causal role belongs to pain with analytical necessity—"by definition"—as the definite characteristic of being in pain. Yet it is contended that in fact, the causal role belongs to some physical state: that physical state *occupies* precisely the role.[27] As empirical science advances, it can confirm that the occupant of this causal role is identical (at least within the relevant kind or species) to a certain physical state. There may be, and likely are, intermediate steps in lower-level special sciences, such as neuroscience, physiology, biology, or chemistry. But in as much as *their* relevant concepts are themselves in turn defined in terms of causal roles in *their* relevant theories, and are for their part realized by states at a lower level, we should presumably be able to repeat the procedure and continue until we are at the fundamental physical level. At least, this would appear to be a natural consequence of the general picture of the concepts of the Canberra Plan, if it is followed consistently.

The presented picture is certainly *prima facie* appealing, but does it bear closer inspection? I have several worries. To begin with, it is not obvious that there is, in general, any sufficiently stable and shared set of platitudes, or a folk theory, and a related causal role. Furthermore, even if the community collectively possessed a sufficient set of "platitudes," it is not clear that a particular member of the community, or even a majority, would know all the relevant platitudes.[28] Note that the ignorant individual could well be our philosopher attempting to follow the Canberra Plan. This might perhaps be a problem for some interpretations of the Canberra Plan. However, I will set this issue aside, go along with the program, and assume that there are clear platitudes and causal roles available.

My key concern is this: The whole approach seems to tacitly assume that the relevant *causal role*, though first defined in terms of the folk theory and common

[26] Of course, this particular analysis is debatable, and it could be argued that it ignores in particular the qualitative feel aspect of pain. I am emphatically not committing myself to this analysis; I only use it to illustrate the program; after all, this was a standard example of analytic functionalism in the 1960s.

[27] At least in Lewis, the assumption of the causal closure of the physical (under the label "the explanatory adequacy of physics") is in play here.

[28] I am not thinking here only of some outré members of the community; I am inclined to think that there can be a widespread division of epistemic and linguistic labor in a developed community. According to Nolan (2010), the platitudes in the Canberra Plan concerning a certain topic "are significant truths about that topic that are implicitly believed by *most*, *or all*, competent speakers" (my emphasis). If so, the above may be a problem.

sense concepts, or of a higher-level special science, can be equally expressed in the vocabulary of physics, and be recognized at that level.[29]

In particular, although it is apparently never explicitly stated by Canberra Planners, it looks as if they tacitly assume something like the following principle, explicitly formulated by Kim[30]:

Causal inheritance principle (CIP): If a higher-level property H is realized in a system at time t in virtue of physical realization base P, the causal powers of this instance of H are *identical* to the causal powers of P (cf. Kim 1993, p. 326).

The essential inference of the Canberra Plan is often presented schematically as follows (cf. e.g., Lewis 1972; Braddon-Mitchell and Jackson 1996, p. 92; Kim 2006, p. 280):

(1) The higher-level state H = the occupant of the causal role R (conceptual analysis)
(2) The occupant of the causal role R = the lower-level state P (empirical discovery)
(3) $H = P$ (transitivity of identity)

The main focus of Canberra Plan has been on (1), but the additional claim, typically considered only in passing, is that the advance in empirical science provides premise (2). Unfortunately, very little is said about how more exactly this is supposed to proceed. It seems that something like CIP might be at work here. That is, there appears to be some kind of unnoticed gap (see below), and it is not clear how it is supposed to be filled. CIP would at least account for that. In any case, the conclusion (3) should then follow trivially.

Now the standard statements of Canberra Plan suggest that the program would allow one, with definitions of theoretical terms, to derive the more theoretical truths, for example those of microphysics, from more familiar O-truths. Yet we should recall that the further idea of at least Jackson and Chalmers (and Lewis) is that all this also

[29] It is worth noting that analytic functionalism (as an early, specific version of the Canberra Plan) was first formulated in the works of Lewis (1966) and Armstrong (1968), before Lewis had developed his counterfactual theory of causation (Lewis 1973). Perhaps there were some other intuitions about causation tacitly in play that suggested (see below) that the causal role can be unproblematically recognized at a lower level. For example, if causation is understood in accordance with the regularity theory of causation (which was still influential in the 1960s), it may appear plausible that the laws of fundamental physics, when applied to successive microphysical states, directly amount to causation. But as Lewis, among others, later argued, the regularity theory is just untenable.

Strikingly, such leading Canberra Planners as Jackson and Pettit (1988, 1990) make a distinction between "causal efficacy," which is causation in the full-blooded sense, and the weaker "causal relevance," which applies to higher-level special sciences and their explanations, and is (in the words of Loewer 2002) mere "causation lite". They argue that genuine causation—efficacy—occurs only at the fundamental physical level. According to them, higher-level states and properties may nevertheless be used in explanations, and are in this sense (causally) relevant. It is unclear to me how the crucial notion of the causal role should be interpreted in this framework: "causal roles" in (higher-level) special sciences and folk theories do not seem to involve proper causation; genuine causal efficacy, on the other hand, is not generally recognized by common folk.

Consequently, I prefer to stick to the more standard counterfactual approach to causation here. (More precisely, my own view is that the interventionist theory, developed especially by Woodward (2003)—which is a more sophisticated counter factualist theory—is the most promising approach to causation available nowadays. From its perspective, higher-level causation is genuine causation as much as anything else.).

[30] Although not usually counted as a Canberra Planner, Kim has also been deeply influenced by Lewis, and many of Kim's ideas can be viewed as elaborations and further developments of Lewis' ideas.

provides, in reverse, an *a priori* entailment of *all* ordinary higher-level truths[31] from the truths of physics (Jackson 1994a, b, 1998; Chalmers 1996, 2012; Chalmers and Jackson 2001). Put differently, in the terminology of Chalmers (2012), all the truths of the higher-level sciences are *scrutable* from the truths of physics.

Be that as it may, it seems to be a part of the Canberra Plan that identity (2) is not only an empirical discovery, but also that it is established solely by lower-level science, and in the end, by fundamental physics. Yet how could it ever entail (2) without talking about the (essentially higher-level) causal role *R*? Presumably, the language of the more fundamental science must somehow be able to talk about *R* too. Therefore, it seems necessary that lower-level science somehow contains this notion as well. Only then can it entail *a priori* identity (2). Although it has gone long unnoticed, there is a mysterious leap here in the program.

However, it is hardly evident that such causal roles can in general be expressed in the language of fundamental physics. Apparently, they are often instead formulated essentially in the language of some (non-reducible) higher-level special science, or folk concepts. At least on one occasion, Lewis explicitly requires that the causal role is recognized by common folk (Lewis 1994).

It may be tempting to think that causation is just something physical, and therefore we can always recognize the causal roles at stake at the level of physics.[32] I am inclined to agree that causation is indeed physical—at least in the broad sense of "physical": there are, probably, no immaterial souls and such, or radically non-physical causal relations between them. Furthermore, it is a plausible assumption that everything—at least everything capable of standing in a causal relation, and causal relations themselves—supervenes on the facts of fundamental physics.[33] Nevertheless, it is a mistake to assume that causal roles identified at some higher level must therefore have exact type-to-type counterparts in the language of physics, and could be recognized at that level. I contend that one may well find no such things at that level.

To begin with, many competent philosophers contend that the concept of causation does not even make clear sense at the level of fundamental physics—that there is no causation at that level.[34] They argue that the concept of causation rather has its home in the higher-level special sciences, e.g. the biological and medical sciences. Be that as it may, our common notion of causation, and the notion used in the special sciences, is essentially a higher-level notion. Even if it were possible to define some sort of notion of causation at the level of fundamental physics, it would likely lack many

[31] Except perhaps, for Chalmers, for example, those that deal with phenomenal consciousness.

[32] The following early statement of the program by Lewis on the case of experience may suggest such a view: "My argument is this: The definitive characteristic of any (sort of) experience as such is its causal role, its syndrome of most typical causes and effects. But we materialists believe that these causal roles which belong by analytic necessity to experiences belong in fact to certain physical states. Since those physical states possess the definitive characteristics of experience, they must be the experiences." (Lewis 1966, p. 17).

[33] Phenomenal consciousness or qualia are, of course, sometimes thought to contradict the assumption. However, we can bracket that issue in this instance: we can assume here that the higher-level facts at stake are objective facts and do not essentially involve qualia or such.

[34] See, e.g., Latham (1987), Redhead (1990), Field (2003); cf. Loewer (2002), Hitchcock (2007), Elga (2007); this idea goes back, of course, to Russell (1912–1913). Vaassen (2020) contains an excellent summary.

characteristics of the more common notion of causation. Moreover, inasmuch as this is the case, the causal roles found in folk theories and higher-level special sciences simply do not have type-type counterparts in the language of fundamental physics.

Furthermore, many favorite examples of Canberra Planners, such as "Water is H_2O," are paradigms of type-identity. However, it is far from uncontroversial that all higher-level properties are as neatly reducible. Even many Canberra Planners seem to allow that in many inter-level cases, no such type-identities exist, and that at least some higher-level properties are multiply realizable.

The informal Canberra Plan revolves around concepts that are (allegedly) *definable* in terms of causal roles. I contend that such properties in particular tend to be regularly multiply realizable. Consider, for example, the concept of *poison*. A rough, simplified analysis of it might be something like[35]:

> A substance *X* that when introduced into an organism causes that organism to die.

Even if we focus only on humans, this property is quite clearly multiply realizable—that is, there are numerous different substances that could realize this role: arsenic, botulinum, cyanide, polonium, strychnine, etc. We can imagine a scenario in which, say, cyanide happened to be in practice the only poison, at least on Earth. It is then conceivable that the scientists would declare that they have "reduced" the property of being poison to that of being cyanide.[36] However, would the two properties be really identical in such a scenario? I submit that they are not: if another realizer is even metaphysically possible, this is enough to make the two properties non-identical, but I digress.

Let us now reflect upon how the basic ideas of the informal Canberra Plan and the *a priori* entailment thesis (of Jackson and Chalmers) work if the relevant concepts are multiply realizable.[37] Let us consider the familiar setting, also dear to many Canberra Planners, of mental states such as beliefs and desires.[38] (Let us again put off phenomenal consciousness, qualia, etc., which may require a different story.) Are truths about such mental states *a priori* scrutable from physical truths?

[35] The example is borrowed (though further simplified) from another central figure of analytic functionalism, Armstrong (1981). Again, I am by no means suggesting that this analysis is fully adequate.

[36] I am emphatically not suggesting that scientists would thereby do something wrong; I am only inclined to think that their understanding of "reduction" tends to be more flexible and contextual compared to the standard held by philosophers for genuine property identity.

[37] Of course, there already exist various notable critical discussions of the *a priori* entailment thesis: at least Block and Stalnaker (1999), Byrne (1999), Diaz-Leon (2011) and Elpidorou (2014) deserve to be mentioned. They make many apt critical observations with which I sympathize, but I don't think that any of them put their finger on exactly the same issue I am discussing here. Further, my interest here is mainly only in the relevance of the Canberra Plan and its tools for the *a priori* entailment thesis, not the varied debate surrounding the thesis outright. It seems that none of these critics has really paid much attention to this particular aspect of the *a priori* entailment thesis. In this respect, I think my analysis here reinforces these critical discussions.

[38] I am focusing here on mental properties only because of their familiarity: as I have noted, the informal Canberra Plan is in many ways just a generalization of analytic functionalism in the philosophy of mind. However, I am inclined to think that the situation would be quite similar with many biological properties, for example.

According to the common functionalist story, also popular among Canberra Planners, these mental states are (in a sense) "definable" in terms of perceptual input, behavioral output, other mental states, and causal relations between them.[39] Perhaps Canberra Planners tacitly assume that at least the defining properties in the causal role, for example, the perceptual input and the behavioral output, in the case of a mental property, are un-problematically physical. However, I contend that even this is not in general true if "physical" here means that the property is expressible in the language of physics.

Consider, for example, the following case: Assume John desires to please his boss, and he believes the best way to do so now is to assent publicly to the boss's proposal to the board. John also observes that the boss attends. The desire, the belief, and the observation together then cause John to assent publicly to the proposal. (Let us assume that John's desire to please the boss is his only reason to assent to the proposal, and that this desire is not overridden by other desires, etc.)

We may assume that the corresponding counterfactual conditional,

> If John had not desired to please his boss, he would not have assented to the boss's proposal

is then true, and that hence, following the counterfactual account of causation, the desire can be taken to be a cause of John's action. Let us assume that this desire is at least partially defined by its causal role—that is, its disposition to cause, in the presence of the belief that the best way to please the boss is to assent publicly to the boss's proposal to the board, and the observation that the boss attends, assent to the proposal.

The perceptual stimulus is John's observation that the boss attends, but the latter seems to be multiply realizable: the boss may be physically present, he might be online in a teleconference meeting, or they may be together in a chat, etc. Similarly, the behavioral response, namely, assenting, is multiply realizable[40]—there are multiple ways to assent—as is presumably the belief that the best way to please the boss now is to assent publicly to the boss's proposal to the board.

In other words, not only the desire at stake but also its defining properties are essentially higher-level phenomena and plausibly themselves multiply realizable. They are emphatically not something that can be found in the language of physics.

Consequently, it just does not seem plausible that we could, only by reflecting on various scenarios presented in terms of physics, tell *a priori* whether the relevant counterfactual claims about beliefs, desires, and assent hold in them or not, or that we could infer the causal role of such mental states from the truths of physics.[41]

[39] I am not myself sure whether all beliefs and desires can actually be "defined" purely in terms of their causal roles, or whether their representational content must be taken as independently given—representational contents, though, seem to be just as much multiply realizable as causal roles, so this is not a relevant issue as regards the question of reducibility (in the sense of type-identity).

[40] Cf. Jackson (1996, p. 389).

[41] Chalmers (1996, pp. 79–80) shows some awareness of this problem, but bypasses it quickly, and defers to Lewis and the Ramsey sentence method. However, we have already seen that the latter faces serious problems, and it is not a trouble-free way out.

Perhaps it is now suggested that these properties can be in turn defined in terms of further things, and in the end, we have something that is both purely physical (expressible in the language of physics) and something recognizable by the common folk. However, this is at best a strong hypothesis in need of substantial defense. At worst, it leads to global structuralism that cuts off reality. At the least, an informal version of the Newman objection applies. In any case, it remains a mystery how exactly Canberra Planners think they can recognize this causal role and its realizer at the level of physics. There seems to be a gap in the whole program here.

As for *the causal inheritance principle* (CIP): Let us assume that the desire is multiply realizable. Now contrary to the popular view, the subvening physical property that realizes the desire does *not* necessarily have exactly the same causal profile as the realized mental property.[42] Assume the desire was actually realized by the physical property P_1, but that it could also be realized, for example, by the physical property P_2 (and many others). The relevant counterfactual conditional,

If John had not had P_1, he would not have assented to his boss's proposal

is then false: even if he did not have P_1, but happened to have, for example, P_2, he would still have assented to the proposal. The underlying physical property here, P_1, even if sufficient (given other conditions) for the effect, is arguably not a difference-making cause (cf. Menzies 2008; Raatikainen 2010, 2013; List and Menzies 2009).

Even if it made sense to talk about the causal profile of the physical property P_1, that profile is not identical to that of the realized mental property in a case like this. Furthermore, the causal profile of the realized mental property apparently cannot be *a priori* deduced from the causal profile of the particular underlying physical realizer P_1 (whatever it is).[43]

If the relevant bridge-laws (at least for the defining concepts) were already independently available,[44] the above issues would not be so much a problem, but it has been an important part of Lewis' view, and presumably also of the Canberra Plan, to do without any independently postulated bridge-laws: bridge-laws may not be a part of the basic language of physics, because they would involve higher-level predicates too.

[42] This issue was raised in the author's correspondence with the late Peter Menzies. It is stated in passing in the work of List and Menzies (2009; see also Raatikainen 2013).

[43] A referee suggested that perhaps Shoemaker's idea, according to which the causal powers of a mental property are (not identical with but) a *subset* of the causal powers of the physical property that realizes it (see e.g. Shoemaker 2001), would enable the problem to be circumvented. I must disagree. To begin with, if there is no causation at all at the level of fundamental physics, as many philosophers have concluded (see above), then there will be no set of causal powers to be a subset of. But independently of that, if the preceding argument has any bite at all, it also shows that the mental property may sometimes be a difference-making cause when the more fine-grained physical property that realizes it is not, thus undermining Shoemaker's thesis: the causal powers of a mental property may just be distinct from and not a subset of the causal powers of the realizing physical property. .

[44] I am thinking of one-way bridge laws here. Of course, inasmuch as the higher-order properties are multiply realizable, there are no two-way bridge laws—i.e. equivalences or identities. However, it is consistent with this to assume that there are still one-way bridge laws (and the supervenience assumption entails that there exist such one-way laws; accordingly, Chalmers (1996, p. 53) calls such one-way laws "supervenience conditionals"). Chalmers (2012), for example, seems to take this line. The argument now is about whether these one-way bridge laws are *a priori* knowable or not, given the truths of physics (as Chalmers and Jackson contend).

Rather, the idea is that the Plan itself provides such bridge-laws.[45] However, without the relevant (alleged) bridge-laws at hand, it may be impossible to pair the language of physics and causal roles characterized essentially in terms of higher-level properties (which are arguably often themselves multiply realizable) in the intended way.[46]

The alternative picture that I am inclined to favor is, very roughly, the following: the scientific community pursues various sciences of different levels side by side; a special science often proceeds more or less autonomously. Now and then, in favorable circumstances, scientists may succeed in establishing correlations and even bridge-laws between different levels, but this is largely empirical activity grounded in developing the sciences of different levels concurrently. Moreover, all of this is fallible, and perhaps the true bridge-laws are only achievable together with the hypothetical complete final theories of different levels (and the latter is only a regulative idea, an ideal limit of scientific inquiry, and not something that is ever reached). Consequently, the bridge laws are known, according to this picture (if at all) only *a posteriori*, and cannot, even in principle, be deduced *a priori* from fundamental physics alone.

5 Conclusions

The formal Canberra Plan is essentially founded on the Ramsey–Carnap–Lewis method. However, it runs up against more than one serious technical problem. Anyone who builds an entire philosophical system using this toolbox should at least deal with these issues in detail and argue convincingly that the particular system is not vulnerable to them. Without such a defense, the prospects of the formal Canberra Plan do not look at all promising.[47]

I have argued that the informal Canberra Plan involves an unclear leap from a causal role, defined at a higher level, to the level of physics. Things become especially troublesome when a certain amount of multiple realizability is allowed. At the very least, both the notion of the causal role and how exactly this step is supposed to proceed should be formulated much more exactly than has yet been done in the literature. The

[45] Thus, Lewis writes: "I deny that the bridge laws must be posited independently. They may *follow* from the reducing theory, via the definitions of the theoretical terms of the reduced theory." (Lewis 1970, p. 427; my emphasis) In a later work, Lewis (1972, p. 248) contends that the lower-level theory *implies* the relevant inter-level identities. Lewis himself assumes here the type-identity theory, but it seems that Canberra Planners assume that what he says here about the inter-level identities applies, *mutatis mutandis*, to the one-way bridge laws. And inasmuch as this is the case, the *a priori* entailment thesis, explicitly advocated especially by Jackson and Chalmers, is at least implicitly built into the Plan generally.

[46] Assuming the situation is reflected more formally (whether or not involving Ramsey sentences): Note that without bridge-laws, the two languages, the language of physics and that of a higher-level special science, are fully distinct: they share no nonlogical vocabulary. Then again, in a great many cases, no truths of the higher-level language (except those that are themselves logical truths) can be derived from the truths of the physical language; it is possible, under quite general conditions, to construct a counter-model in which the latter are true but the former are false.

[47] Canberra Planners have mostly ignored these critical issues. Chalmers (2012), though, repeatedly mentions the "Newman problem," and agrees that it is a fatal obstacle for global structuralism. However, he also states repeatedly that the Newman problem can be avoided by leaving some non-logical expressions unramsified (pp. 8, 21, 363, 409). I am afraid, though, that Chalmers greatly underestimates the scope and difficulty of the Newman problem, and apparently fails to understand the contemporary Newman objection.

burden is obviously on the advocate of the Canberra Plan to present a more rigorous and plausible formulation. In its existing sketchy form, the program appears quite unsatisfactory.[48]

Acknowledgements I grateful to Tim Button, Sean Walsh, and Jeff Ketland for their advice over time with Ramsey sentences and related issues, and to Daniel Nolan for his help with interpreting Lewis and the Canberra Plan. I would also like to thank the anonymous referees for their valuable comments and suggestions that greatly improved this paper. Of course, I am still solely responsible for the claims and arguments made in this paper.

References

Ainsworth, P. M. (2009). Newman's objection. *The British Journal for the Philosophy of Science, 60,* 135–171.

Armstrong, D. M. (1968). *A materialist theory of the mind*. London: Routledge.

Armstrong, D. M. (1981). The causal theory of mind. In D. M. Armstrong (Ed.), *The nature of mind and other essays* (pp. 16–31). St. Lucia: University of Queensland Press.

Block, N., & Stalnaker, R. (1999). Conceptual analysis, dualism, and the explanatory gap. *Philosophical Review, 108,* 1–46.

Bohnert, H. G. (1967). Communication by Ramsey-sentence clause. *Philosophy of Science, 34,* 341–347.

Braddon-Mitchell, D., & Jackson, F. (1996). *Philosophy of mind and cognition: An introduction*. Oxford: Wiley.

Braddon-Mitchell, D., & Nola, R. (Eds.). (2009). *Conceptual analysis and philosophical naturalism*. Cambridge MA: MIT Press.

Button, T., & Walsh, S. (2018). *Philosophy and model theory*. Oxford: Oxford University Press.

Byrne, A. (1999). Cosmic hermeneutics. *Philosophical Perspectives, 13,* 347–384.

Carnap, R. (1958). Beobachtungssprache und theoretische sprache. *Dialectica, 12,* 236–248. **English translation: Carnap 1975**.

Carnap, R. (1959). Theoretical concepts in science. In *Lecture delivered at American Philosophical Association, Pacific Division, at Santa Barbara, California, on 29 December 1959*. Edited by S. Psillos. Published as a part of S. Psillos: Rudolf Carnap's 'Theoretical concepts in science'. *Studies in History and Philosophy of Science*, Vol. 31, No. 1, pp. 151–172, 2000.

Carnap, R. (1963). Replies and expositions. Carl G. Hempel on scientific theories. In P. A. Schilpp (Ed.), *The philosophy of Rudolf Carnap. The library of living philosophers* (Vol. 11, pp. 958–966). La Salle: Open Court.

Carnap, R. (1966). *Philosophical foundations of physics: An introduction to the philosophy of science.* Ed. by M. Gardner. New York: Basic Books.

Carnap, R. (1975). Observation language and theoretical language. In J. Hintikka (Ed.), *Rudolf Carnap, logical empiricist* (pp. 75–85). Dordrecht: Reidel.

Chalmers, D. (1996). *The conscious mind*. New York and Oxford: O.U.P.

Chalmers, D. (2012). *Constructing the World*. Oxford: O.U.P.

Chalmers, D., & Jackson, F. (2001). Conceptual analysis and reductive explanation. *Philosophical Review, 110,* 315–361.

[48] Many of these critical observations concerning the Canberra Plan were first presented, in a much more condensed form, in my earlier paper (Raatikainen 2014).

Demopoulos, W., & Friedman, M. (1985). Critical notice: Bertrand Russell's *the analysis of matter*: its historical context and contemporary interest. *Philosophy of Science, 52,* 621–639.
Diaz-Leon, E. (2011). Reductive explanation, concepts, and a priori entailment. *Philosophical Studies, 155,* 99–116.
Elga, A. (2007). Isolation and folk physics. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality* (pp. 106–119). Oxford: Oxford University Press.
Elpidorou, A. (2014). Blocking the *a priori* passage. *Acta Analytica, 29*(3), 285–307.
Field, H. (1973). Theory change and the indeterminacy of reference. *Journal of Philosophy, 70,* 462–481.
Field, H. (2003). Causation in a physical world. In M. Loux & D. Zimmerman (Eds.), *Oxford handbook of metaphysics* (pp. 435–460). Oxford: Oxford University Press.
Hempel, C. G. (1958). The theoretician's dilemma: A study in the logic of theory construction. In H. Feigl, et al. (Eds.), *Concepts, theories, and the mind-body problem. Minnesota studies in the philosophy of science* (Vol. 2, pp. 37–98). Minneapolis: University of Minnesota Press.
Hitchcock, C. (2007). What Russell got right. In H. Price & R. Corry (Eds.), *Causation, physics, and the constitution of reality* (pp. 45–65). Oxford: Oxford University Press.
Jackson, F. (1994a). Armchair metaphysics. In J. O'Leary-Hawthorne & M. Michael (Eds.), *Philosophy in mind* (pp. 23–42). Dordrecht: Kluwer Academic Publishers.
Jackson, F. (1994b). Finding the mind in the natural world. In R. Casati, B. Smith, & S. L. White (Eds.), *Philosophy and the cognitive sciences* (pp. 227–249). Vienna: Holder-Pichler-Tempsky.
Jackson, F. (1996). Mental causation. *Mind, 105,* 377–413.
Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Clarendon Press.
Jackson, F., & Pettit, P. (1988). Functionalism and broad content. *Mind, 97,* 381–400.
Jackson, F., & Pettit, P. (1990). Program explanation: a general perspective. *Analysis, 50,* 107–117.
Ketland, J. (2004). Empirical adequacy and Ramsification. *The British Journal for the Philosophy of Science, 55,* 409–424.
Ketland, J. (2009). Empirical adequacy and Ramsification II. In A. Hieke & H. Leitgeb (Eds.), *Reduction, abstraction, analysis: Proceedings of the 31st international ludwig wittgenstein symposium in Kirchberg, 2008* (pp. 29–46). Lancaster: Gazelle Books.
Kim, J. (1993). *Supervenience and mind: Selected philosophical essays*. Cambridge: Cambridge University Press.
Kim, J. (2006). *Philosophy of mind* (2nd ed.). Boulder: Westview Press.
Latham, N. (1987). Singular causal statements and strict deterministic laws. *Pacific Philosophical Quarterly, 68,* 29–43.
Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy*, *63*, 17–25.
Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy, 67,* 427–446.
Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50,* 249–258. **Reprinted in Lewis 1999. (Page references are to the reprint)**.
Lewis, D. (1973). Causation. *The Journal of Philosophy, 70,* 556–567.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61,* 343–377.
Lewis, D. (1984). Putnam's paradox. *Australasian Journal of Philosophy, 62,* 221–236. **Reprinted in Lewis 1999. (Page references are to the reprint)**.
Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to philosophy of mind* (pp. 412–431). Oxford: Blackwell Publishers. **Reprinted in Lewis 1999. (Page references are to the reprint)**.
Lewis, D. (1997). Naming the colours. *Australasian Journal of Philosophy, 75,* 325–342. **Reprinted in Lewis 1999. (Page references are to the reprint)**.
Lewis, D. (1999). *Papers in metaphysics and epistemology*. Cambridge: Cambridge University Press.
List, C., & Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106,* 475–502.
Loewer, B. (2002). Comments on Jaegwon Kim's *mind and the physical world*. *Philosophy and Phenomenological Research, 65,* 655–662.
Maxwell, G. (1970). Structural Realism and the Meaning of Theoretical Terms. In *Minnesota Studies in the Philosophy of Science* 4 (pp. 181–192). Minneapolis: University of Minnesota Press.
Menzies, P. (2008). Exclusion problem, the determination relation, and contrastive causation. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced* (pp. 196–217). Oxford: Oxford University Press.
Menzies, P., & Price, H. (2009). Is semantics in the plan? In *Braddon-Mitchell & Nola 2009* (pp. 159–182).
Newman, M. H. A. (1928). Mr. Russell's causal theory of perception. *Mind, 37,* 137–148.

Niiniluoto, I. (1972). Inductive systematization: Definition and a critical survey. *Synthese, 25,* 25–81.
Niiniluoto, I. (1973). Empirically trivial theories and inductive systematization. In R. Bogdan & I. Niiniluoto (Eds.), *Logic, language and probability: A selection of papers from the IVth international congress in logic, methodology and the philosophy of science* (pp. 108–114). Dordrecht: Reidel.
Nolan, D. (2009). Platitudes and metaphysics. In *Braddon-Mitchell and Nola 2009* (pp. 267–300).
Nolan, D. (2010). The Canberra plan. In G. Oppy & N. N. Trakakis (Eds.), *Companion to philosophy in Australia and New Zealand* (pp. 98–100). Melbourne: Monash University Press.
Nolan, D. (2015). Lewis's philosophical method. In B. Loewer & J. Schaffer (Eds.), *A companion to Lewis* (pp. 25–39). Chichester: Wiley.
O'Leary-Hawthorne, J., & Price, H. (1996). How to stand up for non-cognitivist. *Australasian Journal of Philosophy, 74,* 275–292.
Psillos, S. (1999). *Scientific realism: How science tracks truth*. New York and London: Routledge.
Psillos, S. (2000). Carnap, the Ramsey-sentence and realistic empiricism. *Erkenntnis, 52,* 253–279.
Psillos, S. (2006). Ramsey's Ramsey sentences. In M. C. Galavotti (Ed.), *Cambridge and Vienna: Frank P. Ramsey and the Vienna Circle* (pp. 67–90). Dordrecht: Springer.
Putnam, H. (1962). What theories are not. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology and philosophy of science* (pp. 240–251). Stanford: Stanford University Press.
Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). Pittsburgh: University of Pittsburgh Press.
Raatikainen, P. (2010). Causation, exclusion, and the special sciences. *Erkenntnis, 73,* 349–363.
Raatikainen, P. (2011). On Carnap sentences. *Analysis, 71,* 245–246.
Raatikainen, P. (2012). Ramsification and inductive inference. *Synthese, 187,* 569–577.
Raatikainen, P. (2013). Can the mental be causally efficacious? In K. Talmont-Kaminski & M. Milkowski (Eds.), *Regarding mind, naturally* (pp. 138–166). Cambridge: Cambridge Scholars.
Raatikainen, P. (2014). Chalmers' blueprint of the world. *International Journal of Philosophical Studies, 22,* 113–128.
Ramsey, F. (1929/1931). Theories. In F. Ramsey (Ed.), *The foundations of mathematics and other essays* (pp. 212–236). London: Routledge and Kegan Paul 1931.
Redhead, M. (1990). Explanation. In D. Knowles (Ed.), *Explanation and its limits* (pp. 135–154). Cambridge: Cambridge University Press.
Russell, B. (1912–1913). On the notion of cause. *Proceedings of the Aristotelian Society* 13, 1–26.
Scheffler, I. (1963). *The anatomy of inquiry*. New York: Alfred A. Knopf.
Scheffler, I. (1968). Reflections on the Ramsey method. *Journal of Philosophy, 65,* 269–274.
Shoemaker, S. (2001). Realization and mental causation. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents* (pp. 74–98). Cambridge: Cambridge University Press.
Tuomela, R. (1973). *Theoretical concepts*. Wien: Springer.
Tuomela, R. (1974). Review. *Journal of Symbolic Logic, 39,* 617–619.
Vaassen, Bram. (2020). Causal exclusion without causal sufficiency. *Synthese (forthcoming)*. https://doi.org/10.1007/s11229-020-02723-y.
van Fraassen, B. (1997). Putnam's paradox: Metaphysical realism revamped and evaded. *Philosophical Perspectives, 11,* 17–42.
Woodward, J. (2003). *Making things happen*. Oxford: Oxford University Press.