dialectica

International Journal of Philosophy

Contents

EUROPEAN
SOCIETY
FOR
ANALYTIC
PHILOSOPHY

Vol 74 : 4
December 2020

dialectica

International Journal of Philosophy
Official Organ of the European Society of Analytic Philosophy

founded in 1947 by Gaston Bachelard, Paul Bernays and Ferdinand Gonseth dialectica

vol 74
issue 4

December 2020

Contents

Metalinguistic Monstrosity and Displaced Communications

Graham Stevens

David Kaplan's semantic theory for indexicals yields a distinct logic for indexical languages that generates contingent a priori truths. These special truths of the logic of indexicals include examples like "I am here now," an utterance of which expresses a contingent state of affairs and yet which, according to Kaplan, cannot fail to be true when it is uttered. This claim is threatened by the problem of *displaced communications*: answerphone messages, for example, seem to facilitate true instances of the negation of this supposed logical truth as they allow the agent of the message to no longer be at the location of the message when it is encountered by an audience. Many such displaced communications can be identified in everyday natural language uses of indexicals. Recent discussion has suggested that Kaplan's error is to be overly restrictive in the possible contexts of utterance his semantic theory recognizes, as he fails to acknowledge the possibility of utterances that occur at a context distinct from that in which they are constructed. I reject this diagnosis and defend Kaplan's semantic theory. Displaced communications, I argue, are best understood as resulting from a pragmatically introduced metalinguistic context-shifting operation and hence do not demand revision of Kaplan's semantic theory. I provide an analysis of the pragmatic process underlying this operation and make the case for its merits over those of rival accounts of displaced communications.

David Kaplan's (1989a) semantic theory for indexicals yields a distinctive logic for indexical languages, generating a set of logical truths that are entirely absent from non-indexical languages. These logical truths are notable in that they invalidate the rule of *necessitation* ($\vDash \phi \rightarrow \vDash \Box\phi$), because there are sentences that Kaplan thinks cannot be uttered without being true, despite the fact that they express non-necessary states of affairs. Consequently, they are often cited as examples of contingent a priori truths. Recent discussion of indexicality in the philosophical literature has challenged Kaplan's proposal to grant this privileged status to certain indexical constructions, however, by drawing attention to numerous apparent counter-examples in natural language. These challenges almost unanimously agree that Kaplan is too restrictive in his analysis of the sorts of contexts in which an indexical sentence can be employed.[1] All of the proposed counter-examples appear to show that under certain conditions uses of indexical sentences can align an indexical sentence with a context that is not recognized by Kaplan's theory and, therefore, that Kaplan's apparent cases of contingent a priori truths do not reflect genuine semantic features of English, but only reflect Kaplan's mistaken intuitions about the admissible range of contexts in which indexical sentences may be uttered.

In this paper I will defend Kaplan's semantic theory against this challenge. I will proceed by first arguing that the proposed counter-examples in question are not just the result of aligning an indexical sentence with an unusual context, they are the result of applying a context-shifting operator on the character of an indexical sentence. Kaplan calls an operator on character a "monster" and argues that monsters are entirely absent from the semantics of English. There are, however, metalinguistic devices such as quotation that do behave like monsters, as acknowledged by Kaplan. I will argue that the proposed counter-examples to Kaplan's account all share important similarities with these metalinguistic operators and are thus best understood as resulting from pragmatically introduced metalinguistic operators on constructions that are perfectly acceptable on Kaplan's analysis. I thus conclude that Kaplan's semantic theory does not stand in need of revision to accommodate these examples, and we have every reason to retain the view that indexicals can generate the sorts of contingent a priori truths predicted by Kaplan. I will begin by elucidating Kaplan's theory, then discussing the challenges to that theory. I will next introduce the notion of a monster and present the argument that all of the proposed challenges depend on what I will call *monstrous operations*. I will then defend the view that these particular monstrous operations are best explained pragmatically as resulting from metalinguistic operations, rather than semantic ones. Finally, I will consider some objections and replies.

[1] Some notable exceptions are discussed below.

1 The Logic of Indexicals

Pure indexicals are expressions whose literal meaning both requires a context for saturation and specifies precisely what role context must play in the saturation (in English, examples include “I,” “now,” “today,” some uses of “here,” etc.). Demonstratives differ from pure indexicals insofar as they require an additional demonstration such as a gesture or other such directing intention (in English, examples include “that,” “this,” “she,” some other uses of “here,” etc.). To account for this distinctive class of meanings, Kaplan proposes a two-level semantic theory, coupled with a double-indexing of indexical sentences with formal representations of contextual situations. Firstly, an indexical expression is associated with both a *character* and a *content*. The character of the expression is a function from *context* to content. Less formally expressed, this means that the character can be thought of as a rule governing the contribution required by context in order to fix the semantic value or reference of the term. For example, the first-person English pronoun “I” has a character of the form “the agent of this utterance.” This description specifies a function that will yield a different value depending on who is speaking. Sentences have the same two-level semantic profile and each level is compositionally derived from the individual expressions contained in the sentence. For example, the sentence “I am walking” has a character that maps a context c on to the proposition that a_c is walking, where a_c is the agent of the context c. Different contexts will be mapped to different contents by this function. If character is a function from contexts to contents, then individual agents, objects, times, places and so on will be the contents of individual indexical and demonstrative expressions when they are used in context. The content of an indexical is thus its reference.[2] The content of a sentence is a proposition. It follows that indexical sentences cannot contribute truth-evaluable propositions to a semantic theory without the assistance of context. Hence, within the semantic theory, it is ordered pairs of sentences and contexts that

[2] An insistence that *all* indexicals *must* be directly referential is misplaced if we take the definition of an indexical simply to be that its character is a non-constant function from contexts to contents. There seems no principled reason to exclude quantifiers, predicates, or unarticulated constituents such as those which are regularly posited to fix the comparison class for gradable adjectives, from this list. King’s (2001) quantificational analysis of complex demonstratives, for example, treats such expressions as indexical quantifiers. Similarly Kaplan’s formal language *LD* contains an indexical temporal operator in place of a referring expression as its correlate of “now.” Nonetheless it seems obvious that most paradigm cases of indexicals and demonstratives (“I,” “here,” “this,” etc.) are directly referential.

doi: 10.48106/dial.v74.i4.01

are the bearers of truth-values. Contexts themselves are precisely specified as sequences of parameters corresponding to the indexical elements in the sentence. So, for example, a sentence containing "I," "here," and "now" will demand a context with parameters for an agent, a spatial location, and a temporal location. We also add a world parameter to every context. Thus the context in this case will be of the form $c = \langle a_c, l_c, t_c, w_c \rangle$. The sentence-context pair $\langle s, c \rangle$ models the proposition expressed by the sentence s in the context c. Utterances of sentences are thus indexed to contexts. Similarly, evaluation of propositions requires what Kaplan calls "circumstances of evaluation," which are (at a minimum) pairs of world and times. As we will now see, this double-indexing reveals that some sentences will be true with respect to any context they are paired with, despite expressing propositions which are not true at every circumstance of evaluation (thus, are not necessary).

With the above system outlined, we can make sense of Kaplan's claim to have discovered examples of the contingent a priori in English. Consider the following four sentences:

(a) I am me.
(b) This is of the same chemical kind as that.
(c) I am here now.
(d) I am not alive.

Any utterance of (a) will be true because any context c will be such that $a_c = a_c$. Thus we know a priori that any utterance of (a) will be true. And, as this truth holds in all possible worlds, (a) will be true at every circumstance. Thus (a) is a necessary truth. An utterance of (b) in a context c_1 in which the demonstratum of "this" is a sample of liquid water, and the demonstratum of "that" is a sample of water-ice, will be true. But an utterance of the same sentence with different demonstrata could well be false. Hence it is certainly not true a priori. However, if it is uttered in c_1, then the truth it expresses will hold at every circumstance of evaluation. Thus it is an example of a necessary a posteriori truth. Kaplan also holds that any utterance of (c) must be true, on the grounds that there is no context in which an agent can fail to be at the location of that context at the time of that context when uttering something in that context. All the same, their being located at whatever part of space-time they are at when they make that utterance is obviously just a contingent fact—they could have been elsewhere. Hence the proposition expressed does not hold true at every circumstance. Accordingly (c) is assumed by Kaplan to be a contingent a priori truth. Likewise for (d), no-one can utter this truly (as the saying goes, dead men tell no tales), but there is nothing necessary about one's being alive. Thus an utterance of (d) is known to be false on a priori grounds, but is not *necessarily* false.

What makes Kaplan's defence of the contingent a priori so compelling is that Kaplan's proposed cases require no investment in any kind of metaphysical speculation. They are just immediate consequences of the correct semantic analysis of indexicals. Or so it seemed. Many have reacted to Kaplan's logic of indexicals, however, by pointing out that Kaplan's analysis does not seem to be proceeding on purely semantic grounds but in fact makes significant assumptions about the conditions under which utterances and other forms of communication can be made that are empirically questionable. Consider again the example (c) above. This apparent logical truth is routinely negated as an answerphone recording: "I am not here now. Please leave a message after the tone…". Similarly, one may record a message to be replayed at the reading of one's will that contains (d): "If you are hearing this recording, then I am not alive. I have left you this message to communicate my wishes to you after my death…". In such circumstances it appears that truths, not logical falsehoods as Kaplan's analysis seems to predict, are being communicated.

These counter-examples to Kaplan's analysis highlight, and challenge, an assumption in Kaplan's theory about the interplay between sentence and context. Kaplan's assumption is that contexts of utterance, inscription, or other sorts of linguistic performance, always conform to a minimum structural norm such that agents of utterances are always located in the time and place of that utterance. Indeed Kaplan explicitly acknowledges this, arguing that we must restrict contexts of utterance to these "proper" contexts: "[I]mproper indices are like impossible worlds; no such contexts could exist and thus there is no interest in evaluating the extensions of expressions with respect to them" (1989a, 509). Prima facie, these counter-examples seem to show that Kaplan is wrong. Furthermore, the counter-examples are ubiquitous. Here is another, from Predelli (2005, 43). Jones writes the following note just before leaving his house at 8am, which he then leaves at home for his wife, who is not due to return until 5pm:

> I am not at home now. If you hurry, you'll catch the evening flight to Los Cabos. Meet me in six hours at the Hotel Cabo Real.

In this example, the note is obviously not intended to be, nor will it be, interpreted as indexed to the time at which it was inscribed but, rather, is intended to be indexed to the time at which it will be read. Another example, offered by Corazza, Fish and Gorvett (2002), challenges Kaplan's assumption that the agent, utterer, and referent, of "I" in a given context must always coincide. They invite us to consider the case of an academic who leaves a note on his door saying "I am not here today" to signal his absence when it is read. This, already, is an example equivalent to Predelli's above, but they continue the story by imagining that the academic returns to work several days later and then removes the note and reattaches it to a colleague's door to signal *their* absence. It now seems that the referent of "I" (along with other indexicals) has shifted while its inscriber and possibly even agent have not. What are we to say of these challenges to Kaplan's theory? In the next section I will argue that all of these counter-examples display "monstrous" properties.

2 Monsters and Monstrous Operations

Kaplan gives the name "monster" to any would-be operator on the character of an indexical. He maintains that no such operators exist in English. Take for example, the indexical "I." A monster operating on this expression would effect a context shift such that the reference of the expression shifted on to someone other than the agent of the context of utterance. But no such operation seems possible within the confines of ordinary English. If I say "in some contexts, I am not me," this is simply false (if interpreted literally—there may of course be figurative uses of this sentence which are understood to communicate a truth). I am always identical to myself. Similarly, if I embed the same indexical within a propositional attitude verb, the attitude verb has no impact on the character of the indexical, which immediately takes wide scope with respect to it: an utterance of "John believes that I am happy" communicates the speaker's report that John believes the speaker to be happy. The character of "I" picks out the speaker, regardless of any operators contained in the utterance.

Kaplan does point out, however, that monstrous operations can be created by *metalinguistic* devices. The most common is quotation. Compare the following:

(e) John said that I am happy.
(f) John said "I am happy."

By naming the indexical sentence “I am happy” we can shift the character of that sentence in (f), whereas it is impossible in (e). Kaplan’s position, then, is that the only operators resembling monsters that can be applied to English expressions are metalinguistic operators. To keep this distinction between linguistic and metalinguistic operations intact in what follows, I will use the expression “monster” to denote the sort of lexicalized linguistic operator that Kaplan maintains is absent from English,[3] and the expression “monstrous operator” (hereafter “MO”) to denote any operation, including the metalinguistic operators such as quotation, on character. Thus, according to my usage, every monster is an MO but not every MO is a monster.

3 Displaced Communications

The counter-examples to Kaplan’s proposed truths of the logic of indexicals have been responded to in varying ways, but most of the responses conform to one general strategy. The counter-examples are usually understood as somehow involving a displacement from its point of origin of the information that is communicated. To put it another way, a distinction is drawn between the point at which the communication is encoded and the point at which it is decoded. The alleged flaw in Kaplan’s reasoning has thus been almost unanimously identified as the mistaken assumption that communications occur at the point when (and where) they are encoded (either recorded in the cases like answerphone messages, or inscribed in the case of written notes and messages). By contrast, these counter-examples all seem to be intuitively understood as communicating information at the time when they are decoded. Sidelle (1991, 535) describes the production of an answerphone message as a process of “arranging to make an utterance at a later time, or, if one likes, deferring an utterance.” This interpretation has gone largely unchallenged despite the differences in opinion as to the semantic or pragmatic mechanism by which this proposed procedure is thought to be realized.[4] In what follows

[3] Kaplan only says that monsters are absent from English but he is often interpreted as making the wider claim that they are absent from natural languages generally, a claim challenged by Schlenker (2003), who appeals to empirical data concerning Amharic to support the view that monsters are present in some natural languages. For a detailed discussion of Kaplan on monsters, see Predelli (2014).

[4] It is denied by Stevens (2009). Cohen (2013, 8, fn 8) rightly points out that a major shortcoming of Stevens (2009) is my lack of a positive proposal in place of this interpretation. In particular, no detail is given as to how a pragmatic explanation could explain apparent utterances at a distance, so as to make the idea of a deferred utterance redundant. A proposal like the one I will be offering

I will use the term "Displaced Communication" (hereafter "DC") to denote this proposed act of encoding content for later decoding.[5]

It is rarely, if indeed ever, noted that any such act of encoding content for later communication is itself an MO in the sense that all such cases involve the use of an indexical to communicate a content which is shifted away from the context in which the communication is encoded. For example, the use of "now" to refer to a time other than the time at which it is being used to encode the communication, the use of "I" to refer to someone other than the agent who is encoding the communication, and so on. Furthermore, it is clear that the operation of deferring the communication must involve this monstrous operation. If it did not, then we would not have our counter-examples to Kaplan's proposed analysis. DCs are not monsters in the sense of being discrete lexically encoded operators within the language that introduce MOs, but they are certainly monstrous in this wider sense that includes pragmatic processes and operations (which is clearly encompassed by the definition of an MO provided above).

Two objections may be raised against the above description of displaced communications as being (or being the result of) MOs, which should be addressed before continuing. Firstly, one may point to the fact that displaced communications appear to perform (or result from the performance of) an operation directly on the *context* with which the communicated sentence is paired, rather than the *character* of the sentence. Secondly, one may point to the fact that monsters, as traditionally understood, play a recognisably semantic role by operating directly on lexical items through binding operations, imposing scope relations on them, and so forth. I will reply to each objection in turn.

On Predelli's (2005) interpretation of DCs, the correct context that a sentence must be paired with for accurate semantic evaluation is determined by the intention of the speaker. That I can intend my utterance at time t_1 of the sentence "I am not here now" to be evaluated at a later time t_2 demonstrates, on this view, Kaplan's error in defining proper contexts too narrowly. On

in this paper is going to be required if we are to reject the deferred utterance analysis. Others have offered pragmatic proposals that share some features with my approach, including Connolly (2017) who also diagnoses the counterexamples as involving some form of pretence, although his analysis of how this is effected takes a very different line to mine. Åkerman (2017) also provides an alternative account of how pragmatic processes can be appealed to in our explanation of apparent cases of context-shifting.

[5] I prefer the term "displaced communication" to "deferred utterance" as it is not limited to utterances, but can include any communication of information.

this view, the counter-examples to Kaplan's theory arise because sentences have been paired with contexts that he did not recognize. As such, it does not automatically appear that any MO has been applied. The character, Predelli (2005, 44) insists, does not change at all. Rather, careful inspection of the "preparatory operations" (2005, 58)—a phrase Predelli borrows from Quine to describe the decisions we make about how to regiment particular utterances or inscriptions to make them ready for semantic evaluation—simply reveals that sentences can in fact be paired with a wider selection of contexts than Kaplan anticipated. All of the work that separates Predelli's position from Kaplan's occurs at this pre-semantic stage: "[O]nce the appropriate clause-index pair has been identified, the indexicals proceed with their customary characters, and results of truth-value are obtained on the basis of the usual mechanisms of compositional analysis" (2005, 58). Nonetheless, further reflection makes it clear enough that this position entails the existence of MOs. An MO is an operation which shifts the context of an indexical element or sentence by operating on its character. We can easily modify the examples of displaced communication to make this effect more explicit. For example, one can download an audio file of Arnold Schwarzenegger's famous utterance of "I'll be back" from the *Terminator* movie and set this file to play as an answerphone message. This performs an MO on the original utterance (made by Schwarzenegger), shifting the context to one in which the agent is no longer Arnold's cyborg character from the movie, but a real person who has set up an answerphone message to denote their temporary absence or unavailability to a caller.

A further worry may arise by comparing these sorts of MOs with the operators that Kaplan defines as monsters. A monster is an operator on the character of an expression. One way to identify the presence of a monster would be to look for the observable effects that the monster has on the scope of the expression operated on. For example, if we try to shift the scope of the indexical "now" by using a phrase like "in some contexts," we get the following:

(g) In some contexts, yesterday is now.

The attempt to shift the context fails because "now" and "yesterday" resist embedding under the scope of the operator; i.e. they take wide scope over the operator. On the surface, this seems to be a quite different operation to any present in cases of DCs. The answerphone message "I am not here now," for example, does not display any distinctive impact on the scope of the indexicals contained in it. While this is undeniable, I don't think it counts against the view that an MO is at work in cases of DCs.

For one thing, there are many examples of DCs which are quite naturally understood as imposing a binding operation on an indexical element. Take for example the logo used on merchandise by the Rock Climbing equipment manufacturer DMM. They produce clothing with the following phrase emblazoned on it:

(h) Climb now, work later.

This phrase can be naturally used in conversation in a way that pairs it with a proper Kaplanian context:

> *Speaker A*: I don't know whether I should go climbing now, or after I have finished my essay. What do you think?
> *Speaker B*: Climb now, work later.

When it occurs as a logo on the aforementioned clothing, however, it cannot be interpreted this way. There is not just one time which fixes the referent of "now" and "later" in this case. Rather it expresses something along the lines of "always go climbing before doing your work." In fact, it seems to have much the same logical structure as a puzzling case noted (but not addressed) by Kaplan:

(i) Never put off until tomorrow, what you can do today.

The difference, of course, is that (i) contains a lexical item "never" which quantifies over temporal values, making the binding of "tomorrow" and "today" explicit. But this difference is trivial—it is obvious enough that the quantifier "always" is implicitly present as an unarticulated constituent of (h). Thus we have a case of a DC in which the indexicals are forced to take narrow scope with respect to an operator. This is a clear example of an MO.

4 Metalinguistic Monstrosity and Varieties of Quotation

The above sections give us reason to accept that DCs are MOs. I will now argue that, although they are MOs, they are not monsters. This leads naturally to the conclusion that their context-shifting powers are not the result of any semantic operation but is best thought of as the result of a pragmatic process. This will lead me to conclude that DCs pose no serious challenge to Kaplan's theory of indexicals and, in particular, to his restriction of the range of admissible contexts to proper contexts. Consequently, I shall argue that Kaplan is correct to classify sentences like "I am here now" as encoding contingent a priori truths.

That DCs are not monsters is fairly self-evident. Monsters are linguistic operators. No lexicalised operator can be discerned in the DCs discussed in the literature. It is not the addition of a new constituent to the sentence "I am here now" which performs the role of an MO when this sentence occurs within a DC. It is the particular use the sentence is put to. One could perhaps pursue the line that an unarticulated constituent is responsible for the MO and is thus a monster but there seems little evidence or motivation for such a view.[6]

If DCs are MOs but are not monsters, then they are most naturally understood as behaving exactly like the paradigmatic metalinguistic MOs recognized by Kaplan which were discussed earlier. Quotation, for example, is an MO because it can take an indexical sentence and block the indexicals within it from taking their customary wide scope positions. It does so by *mentioning*, rather than *using*, the sentence. I suggest that DCs are the result of MOs which do exactly the same thing. A DC is created by taking an indexical sentence and recording it ready to be mentioned in a new setting at a later date. Construed in this way, DCs are not constructed by a semantic operation. Thus the criticism levelled at Kaplan which draws on DCs as apparent counter-examples to his semantic theory for indexicals is misplaced. A distinctive feature of DCs is that they require a rich contextual setting. This contextual setting is not the minimalistic sequence of parameters required for saturating indexical expressions, but a far wider notion of "context," incorporating various complex conventions surrounding human interaction. These conventions are essential to the performance of a DC. But such features are not semantic features. Thus it is natural to understand the MOs involved in generating DCs as pragmatically licensed, rather than semantic. I will now explain in detail the pragmatic process that I suggest is at work in these cases.

Quotation is surprisingly varied. Whereas it was once assumed that quotation is a simple device for self-nominalisation, enclosing a string of expressions within quotation marks to generate a name of that string, it is now widely

[6] This point is argued for convincingly by Predelli (1996).

noted that quotation is not restricted to this simple operation. Consider a case where it does seem to behave in such a way. For example:

(j) "Schnee" is a German expression which stands for snow.

In (j) the name "Schnee" names the expression "Schnee" but that expression is not used in any meaningful way in (j). We are simply exploiting the convention whereby quotation marks name the expressions enclosed within them. We could just as easily have exploited a different naming convention or indeed stipulated one. For example, I can stipulate a convention for naming an expression as follows:

(k) Let whichever German word stands for snow be called "Angelika." Angelika is often uttered by German speakers when in the presence of snow.

The convention employed in (k) is perfectly clear and comprehensible. It does not require any grasp of the German expression for snow or even an ability to recognize that expression. We simply report facts about the expression by utilising a (descriptively introduced) name of it. These sorts of examples demonstrate that quotation behaves as a purely "mentioning" device in such contexts. The semantic content of the expression that is being quoted is wholly inert in these contexts (hence its unproblematic absence in (k)).

Other uses of quotation, however, are less simple. Newspaper headlines, for example, commonly employ quotation not only to report speech but also to convey information encoded by that speech. Here are a few examples taken at random from the BBC News website on one visit:

> Woman "killed dad and buried him"
> Army ads "won't appeal to new soldiers"
> Financial services "pivotal to Brexit deal"
> "Chronic" nurse shortage and Meghan "mania"

In all of the above cases, the quoted material is a speech report. However, there is more going on here than *just* a speech report. Compare it to the following, more straightforward speech report, taken from Rap Artist Chuck D's (1998, 193) autobiography:

> Ice-T was in the video because I saw him while he was in Arizona and asked him if he wanted to be in the video. He said, 'Cool.'

This is simply a speech report—reporting the words used to accept an offer to feature in a promotional video. In the previous examples, however, the direct quotation is not simply a report of the words used by whoever uttered them—it also draws attention to the speech act they were used for and, in doing so, draws attention to (and *uses them to express*) their content. In this usage, which is very common in newspaper headlines, the quotation does not just name the expressions used but also establishes that they were used to *allege* something. In such cases where an allegation is reported, the reporter does not implicate herself as one making the allegation, she simply reports the allegation and reports that such an allegation has been made. But we can also find cases where quotation is employed not just to report a claim, but also to *endorse* that claim:

(1) Kaplan's example of a kidnapped heiress, locked in the trunk of a car, who has lost all track of time and of her location, yet who can still think to herself "it is quiet here now," demonstrates clearly that, "ignorance of the referent does not defeat the directly referential character of indexicals."

In (1) the quoted material (from Kaplan 1989a, 536) at the end of the passage both reports Kaplan's view and, at the same time, endorses it. The quoted material is not simply named; it's content is asserted. Quotation of this sort, labelled "mixed quotation" by Cappelen and LePore (1997) because of its dual role as reported and asserted, is quite common.[7] Récanati (2010) helpfully distinguishes between "closed" and "open" forms of quotation to make sense of the distinctions at play in these cases. Following Davidson (1979), Récanati interprets quotation marks as performing a demonstrative role whereby the linguistic material (L) quoted is demonstrated as the referent of the quoted expression "L" in reports of the form "S said that 'L'." The difference between open and closed quotation is that closed quotation recruits the demonstrated material to play the syntactic role of a singular term, whereas open quotation is any form of quotation that cannot be so construed:

> The contrast between open and closed quotation is illustrated by the following pair of sentences:

[7] Davidson (1979, 29) first drew attention to such "mixed case[s] of use and mention."

doi: 10.48106/dial.v74.i4.01

> (7) Stop that John! ‘Nobody likes me’, ‘I am miserable’ … Don’t you think you exaggerate a bit?
> (8) John keeps crying and saying ‘Nobody likes me’.
>
> In (7) a token of ‘Nobody likes me’ and ‘I am miserable’ is displayed for demonstrative purposes, but is not used as a singular term, in contrast to what happens in (8), where the quotation serves as a singular term to complete the sentence ’John keeps crying and saying ___’. Sentence (7), therefore, is an instance of open quotation, while (8) is an instance of closed quotation. (Récanati 2010, 231)

If we acknowledge this distinction, we ought to agree with Récanati that there is a fundamental difference in linguistic role between the demonstrated linguistic material in closed quotation and that in open quotation. Whereas closed quotation recruits (a token of) the linguistic material as a singular term which is naturally understood as referring to itself (as a type in most instances), open quotation does not feature any singular term which naturally presents itself as requiring a referential interpretation. Récanati’s proposal is that the sense in which the quoted material acts as a *demonstration* in open quotation is wider than the customary sense in which demonstratives are taken to have their reference fixed by an accompanying demonstration. In open quotation, quoted material *demonstrates* in the sense of providing a performance or picture that represents through a form of mimicry. Thus, in Récanati’s example (7) above, John’s speech is quoted as a means of mimicking his self-indulgent utterances.

Understood as mimicry, open quotation has monstrous potential. This is unsurprising, of course: we have already noted that quotation is a metalinguistic monster. Mimicry is clearly an attempt to represent a content which is, in some sense, displayed from a perspective other than that of the speaker, namely the perspective of the one that the speaker mimics. Consider my report of my teenage daughter’s recent request for a new pair of trainers:

(m) Amy has asked me to get her a “sick” new pair of trainers.

I do not, ordinarily, use the word “sick” with the sense it has been assigned in (m). If I am honest, I confess that I am not entirely sure what the extension of the term “sick” is when my daughter and her friends use it in their conversations. But I know enough about it’s meaning to know that it is a desirable property of footwear in my daughter's opinion (hence this use of the expression is not synonymous with other, more common, uses of "sick" in English) and this is readily communicated (to others who have at least the same level of acquaintance with this term as I do) by my utterance of (m). But I do not communicate *my* judgement as to the possession of this property by any footwear when I use this term in (m); I communicate my daughter's judgement. Thus we have a form of context-shifting operator present in (m). The quoted expression "sick" shifts the context to one in which Amy judges things to have the property that she takes that expression to encode. As Récanati (2010, 260) notes this may not amount to a full-blown MO as it is not clear that shifting from my idiolect to Amy's is best represented by a shift in the sequence of parameters we would normally associate with a linguistic (as opposed to metalinguistic) context. Nonetheless, it illustrates nicely the potential that open quotation has for shifting the perspective away from that of the speaker in a way that is commonplace in ordinary discourse. Indeed we can extend the usual notion of a linguistic context to accommodate such metalinguistic operations easily enough, by incorporating an "idiolect" parameter for the context (see Récanati 2010, 260), according to which my utterance of (m) will be interpreted as employing a context-shifting operator (quotation marks) to recruit the content of the expression "sick" as assigned in Amy's idiolect to act as a constituent of a proposition whose content is otherwise assigned in accordance with my idiolect.

Open quotation, understood thus, has a number of advantages, most notable among them being the fact that we can simultaneously maintain that quotation is (i) metalinguistic, (ii) an MO, and (iii) *used* (rather than merely mentioned) as a means of communicating information. Furthermore, as I will now illustrate, it provides a perfect explanatory model for the MOs discussed in this paper. My proposal is that DCs are best analysed as akin to instances of open quotation. It follows that DCs are MOs but this poses no threat to the Kaplanian claims that there are no linguistic monsters in English, only metalinguistic MOs, and that DCs do not provide counter-examples to the proposal that there are special logical truths of indexicals languages or to any of Kaplan's proposed logical truths of indexical languages.

To see how MOs can be interpreted on the same lines as instances of open quotation, it will be helpful to consider a range of similar phenomena involving intuitive context-shifting operations.

doi: 10.48106/dial.v74.i4.01

Expressives are expressions which encode a speaker-attitude alongside, but independently of, their truth-conditional content (if they have any).[8] If I utter the expression "yummy," upon encountering a delicious foodstuff, I express my positive attitude towards its flavour. But now consider the phenomenon of child-directed speech. In the years before my daughter was old enough to desire "sick" new trainers, I used to feed her baby food: pureed vegetables, rusks, and other assorted foodstuffs which I personally do not find even remotely appetising. Yet, it was common for me to feed her and to accompany the process with enthusiastic utterances of "yummy." Intuitively, I was not expressing my positive attitude towards the taste of the food; rather I was expressing (or perhaps encouraging) my daughter's positive attitude. A simple explanation of what is happening here is that the expressive "yummy" encodes the attitude of the speaker. But in cases like this, we have an implicit open quotation operator, which effects a context shift from speaker-attitude to the attitude of the quoted speaker. This is realised, as in the cases considered above, by an act of mimicry. By mimicking the reaction (or, perhaps, *desired* reaction in this case) of my daughter, I implicitly apply a form of open quotation to her utterances.

We encounter the same thing when we consider expressives with a truth-conditional component. The expression "eurocrat" is a mild, and slightly comical,[9] pejorative expression used by anti-EU British politicians (and those who support them in this regard) to denote the politicians and their fellow officials who form the European Parliament and administer the bureaucracy of the European Union. An utterance of the sentence, "I think it is hilarious that Farage has to spend his time hanging out with all those eurocrats," made by someone who obviously does not share Farage's attitude of contempt towards the bureaucrats in question, is naturally read as an open quotation which would most perspicuously be represented as such:

(n) I think it is hilarious that Farage has to spend his time hanging out with all those "eurocrats."

[8] "Pure" expressives like "ouch" and "oops" appear to make no truth-conditional contribution to utterances and to simply encode a speaker attitude, whereas, e.g., pejoratives (including racial or sexual pejoratives) are often thought to encode both a truth-conditional content (an extension, namely those who the speaker intends to denote by the term) and a speaker attitude (of derogation towards the members of that extension).

[9] Admittedly, its comedic quality has been somewhat diminished since the UK's recent referendum result.

Interpreted thus, it is obvious that the pejorative force of the expression "eurocrat" has shifted away from the speaker-attitude to the attitude of Farage. The use of the term is again interpreted as mimicry.

Predicates of personal taste have also been noted as displaying similar behaviour. If Mary says "Rollercoasters are fun" and John says "Rollercoasters are not fun," Mary and John are disagreeing *faultlessly*. That is to say that although it appears that one is asserting a proposition while one asserts the negation of that proposition, there is a sense in which both are speaking the truth, without either speaker misrepresenting the facts. Semantic Relativists like Lasersohn (2017) explain this by taking the truth of their utterances to be sensitive to a contextual parameter included in the circumstance of evaluation which ensures that the standard for truthful attribution of fun may differ between the two utterances. But now consider the case where Mary asks John, immediately following his rollercoaster ride, "Was that fun?". Intuitively, the relevant standard here is not Mary's but John's. She is asking if *he* found it fun. Again, we can make sense of this scenario by understanding the evaluation of the attribution of the property encoded by "fun" to be relative to a parameter which is usually set to the speaker of the expression but in cases like this is shifted to the addressee. Again this can be understood as resulting from an implicit open-quotation device to mimic the addressee of the question. Note that mimicry here does not have to be a convincing performance, it simply needs to present the attitude or perspective of the target agent to whom the attitude is being attributed. I can do the same thing when I feed my pet guinea pigs some dried pellets of food and ask them, "Is that tasty?". I do not need to be providing a convincing impression of a guinea pig to make it clear that the relevant standard of, and perspective on, tastiness here is that of my guinea pigs (or, at least, that which I attribute to them), not mine.

The above examples demonstrate that context-shifting is familiar for a range of expressions.[10] What then of the content of indexicals? Can we provide

[10] Of course, not all will share my analysis of these cases as instances of open quotation. For example, irony of the sort displayed in examples like (m) and (n) may inspire competing analyses from Griceans. I do not have space here to mount a detailed defence of my analysis of irony and related phenomena, but hope to have made it clear that the analysis is a plausible one for a range of phenomena that are importantly similar to the cases we are concerned with. As well as drawing on Récanati's approach, my analysis has some similarities with the echoic analysis of irony and related phenomena adopted in Relevance Theory (see, e.g., Wilson 2006). The Relevance Theoretic approach is applied by Bianchi (2014) to echoic uses of slurs. An important point, emphasized by her, is that when we echo or imitate the perspectives of others, we do not have to extend the echoing to the whole content of an asserted proposition—we often only express a

doi: 10.48106/dial.v74.i4.01

examples where the same context-shifting operation shifts the reference of an indexical in the way that we expect MOs to do? In fact such examples are easy enough to find. First consider another example of child-directed speech. A nursery teacher, talking to a very young child who had her birthday the day before says: “Did mummy and daddy take you somewhere nice for your birthday?” Not only is the contextual standard for “nice” shifted to that of the addressee, but also the content of the terms “mummy” and “daddy” have shifted. These expressions behave very much like indexicals in that when uttered without qualification, they refer to the parents of the speaker. But here, the only qualification arises as a consequence of the nature of the context. That context generates a construction best understood as an open quotation: “Did ‘mummy’ and ‘daddy’ take you somewhere ‘nice’ for your birthday?” where the open quotation operation shifts the context away from that of the speaker to the addressee in order to fix the content of the quoted expressions.

Of course, it might be replied that this example can equally be explained by appeal to ellipsis. It might be thought that the indexical-like features of “mummy” and “daddy” are best explained by appeal to an elided possessive marker “α’s mummy,” which may be an obviously indexical possessive pronoun “my mummy,” “your mummy,” “her mummy,” etc. Be that as it may, there are other cases which make it perfectly clear that indexicals can be shifted by open quotation. Indeed we saw one above from Récanati, which I repeat here:

(o) Stop that John! “Nobody likes me,” “I am miserable” … Don’t you think you exaggerate a bit?

It is clear that the referent of the “me” and “I” in (o) is not the speaker of (o) but the person they are mimicking, namely John. Examples like (o) are not uncommon and are a clear example of the use of open quotation as an MO that shifts the context that the indexicals contained within it are indexed

perspective distinct from our own with regard to a *part* of that proposition. For example, Bianchi interprets the sentence “As I reached the bank at closing time, the bank clerk helpfully shut the door in my face” as containing an instance of echo or imitation only with regard to the expression “helpfully” (2014, 39). This is the same feature that I am appealing to open quotation to explain in many of the examples above. Bianchi draws on this analysis to explain seemingly non-offensive uses of slurs, such as we see in appropriation (cases where the usual targets of a slur use the expression in a way that removes its derogatory aspect). While I agree that there are echoic uses of slurs such as (n), I would not extend this analysis to appropriation (my own account of appropriation can be found in Scott and Stevens 2019); other examples of echoic uses of slurs and of expressives more generally are discussed in Stevens and Duckett (2019).

away from the parametric settings of the overall sentence to another context for those quoted segments of the utterance. Using subscripts to display the relevant indexes, the situation is something like this:

> [Stop that John! "[Nobody likes me]$_{c_2}$," "[I am miserable]$_{c_2}$" … Don't you think you exaggerate a bit?]$_{c_1}$

John is the addressee of c_1, and not the agent; he is the agent of c_2, not the addressee.

Notice that open quotation is being considered as an explicit (albeit metalinguistic) operation in the above examples. The operator (quotation marks) is ambiguous between open and closed quotation producing functions but it is explicit in the syntax of the written language. But, of course, utterances are not always inscribed. Except in rare cases where quotation marks are "signed" by a gesture which conventionally signals that the words uttered contemporaneously with that gesture are being quoted, it is up to hearers to identify quotation from features of the context. Pragmatic aspects of utterance interpretation come to the fore in such situations. Consider the following pairs of utterances:

(p) That guitarist, John, is performing tonight.
(q) That guitarist, John, who can't actually play the guitar to save his life, is performing tonight.*

The awkwardness of (q) (indicated by the *) follows from the apparent contradiction which results from simultaneously describing John as bearing a property and then denying that he bears that very property. Were one to hear an utterance of (rather than read an inscription of) (q), however, one would most likely apply a principle of charity and interpret the utterance in a way which resolved this potential infelicity, such as (q_1):

(q_1) That "guitarist," John, who can't actually play the guitar to save his life, is performing tonight.

In other words, the term "guitarist" is interpreted as being subject to an open-quotation operator, shifting its usual extension to one that includes John (who is exempt from the extension of the standard English term). It may be read as synonymous with "so-called 'guitarist'," hence behaving much like the term "sick" discussed above: the target of a metalinguistic operator that shifts the idiolect (or other metalinguistic feature) according to which it is interpreted.

This process of pragmatically guided utterance interpretation need not apply only to subsentential elements but can equally be applied to whole sentences. A few years ago, there was something of a craze for purchasing audio recordings to be played as an answerphone message. A popular recording, used as an example above, was the snippet of Arnold Schwarzenegger's character from the *Terminator* movies uttering the line "I'll be back." If I call my friend and hear this message, I do not interpret it as expressing the proposition it was originally used to encode. I interpret it as saying that the person who I have called is temporarily absent and soon to return. It is, in fact, interpreted as if the person being called were able to respond to my call from their current location and say "As Arnold Schwarzenegger says: 'I'll be back!'." In other words, I understand the utterance as displaying the utterance made by Schwarzenegger and recruiting it to communicate information. I interpret it in precisely the same way as an instance of open quotation. The message is interpreted in just the same way as I would interpret the utterance of my friend who explicitly mimicked Schwarzenegger's character, monotone pronunciation (perhaps even accompanied by distinctive bodily movements) and all, when in my presence.

Of course, mimicking an iconic actor or fictional character, by uttering an iconic line from an iconic movie is one thing, but what about ordinary answerphone messages, written notes, etc.? Who, or what, is being mimicked in these cases? Mimicry in these cases is more mundane but best understood as mimicry nonetheless. All that happens in these cases is that the speaker mimics *themselves* saying what they would say, were they able to inhabit the impossible (that is, improper) contexts they would need to be in to otherwise communicate this information. That is to say, when one needs to communicate information from a context unavailable to them, one must find an alternative method of relaying the information. By preparing in advance a message to be retrieved by ones intended audience in this context one is able to overcome this obstacle. But this is achieved not, as is often assumed, by somehow making an utterance "from a distance" but by recording in one form or another an instance of oneself performing the speech act one would want to make at that context if able to, ready to be displayed there. In doing so, one does not encode the proposition that would be obtained by pairing the uttered sentence with the improper context in question, but simply prepares a string of linguistic material that mimics the intended performance and then exploits the various media which permit this mimicry to be planted in advance ready to be deciphered when encountered. What is deciphered is not an utterance by an absent agent, it is a previous utterance deliberately placed in a situation where it will pragmatically trigger a process of interpretation precisely akin to that by which we read utterances of open quotation as discussed above. There is no fundamental difference between my employing open quotation to use an instance of the indexical "I" to pretend to be someone I am not, and my employing open quotation to use an instance of the indexical "here" to pretend to be speaking at a location where I am not. DCs are not utterances made at a distance, they are recorded performances pragmatically recruited to mimic intended and otherwise impossible utterances. This distinction is far from trivial: it demonstrates that DCs are generated by pragmatic features of communication and are thus not data to be accommodated by semantic theory.

The only difference between the uses of open quotation that I have discussed above and the full-blown DCs is that the syntactic role of quotation to mark off the shift will clearly be uncalled for in the latter case. Accordingly, we should not expect quotation to be readily recoverable in a case where a DC consists entirely of a mimicked performance, whereas it is essential for embedded occurrences like we see in "Amy has asked me to get her a 'sick' new pair of trainers." Only when DCs are embedded would explicit quotation marks be felicitous, as we saw with "As Arnold Schwarzenegger would say: 'I'll be back'." Other DCs would need to be placed in similarly embedded constructions to achieve the same result. For example: "If I could speak at the context where you will hear this message I would report that, 'I am not here now'," and so on.

5 Semantics, Pragmatics, and Displaced Communications

In this section I want to briefly say a few things in defence of the view presented above and to point out its advantages over competing accounts of DCs. In recent years the most vocal and influential critic of Kaplan's account of logical truth for indexical languages has been Stefano Predelli. Although couched within a position sympathetic to Kaplan's semantic project, Predelli takes issue with Kaplan's decision to limit the possible combinations of sentences with contexts to proper contexts. Drawing on the DCs discussed above, Predelli argues that Kaplan is simply wrong to assume that utterances require their agents to be located at the times and places they occur. Furthermore, he offers an ingenious proposal as to how an extension of Kaplan's theory to include improper contexts can be motivated and put into practice.

According to Predelli (most comprehensively in 2005) no semantic theory for indexical languages can be complete unless it has the resources to accommodate the role of speaker intentions in fixing the parametric settings of indexicals as uttered. In particular, Predelli maintains that speaker intentions are crucial to determining the context with which an indexical sentence must be paired in order to correctly model the actual utterance. Presented thus, Predelli's position may not sound particularly distinctive—challenges to the attempt to model meaning by formal semantics alone without recognition of the role played by the speaker intentions behind the utterances whose meaning we are attempting to model are common from those who maintain that a theory of pragmatics is needed to explain linguistic meaning. The novelty of Predelli's position however rests on his desire to reconcile his approach with a philosophy of language that assigns the core role of explaining meaning to formal semantics. To bring about this reconciliation of speaker intentions and semantically assigned meanings, Predelli draws a distinction between the workings of a formal system which calculates truth-conditions for utterances and a "pre-semantic" arena in which the inputs to this system must be first determined. It is in this latter area that speaker intentions become significant. Before I can employ a formal system to calculate the truth-conditions of a speaker's utterance, I first must determine *which* utterance she has made (which *proposition*, in other words, she has said). To take a trivial example not involving indexicality, I can only know the truth conditions of an utterance of the sentence "John is sitting beside the bank" if I know which lexeme the ambiguous English word "bank" encodes in that sentence. So I must determine which lexeme the speaker intended before I input her utterance into the formal semantic theory which then returns its truth-conditions as output. The same thing happens when one utters an indexical sentence, according to Predelli, but now the pre-semantic task is to determine the correct context that this sentence must be paired with and, as with the case of disambiguating a lexical ambiguity, the only correct answer here will be that which identifies the speaker's intention.

Certainly there is much to agree with in Predelli's account, and its subtleties are not always recognised by his critics. For example, the objection often directed at Predelli that he is guilty of Humpty-Dumptyism[11] is misguided. Humpty-Dumpty does not resolve ambiguities or select contexts to pair in-

[11] "Humpty-Dumptyism" is the pejorative term for a semantic theory of the bizarre and implausible sort envisaged by Lewis Carroll's fictional character who insisted that his words mean simply whatever he wants them to mean. Responses to the charge are given in Predelli (2011).

dexicals with, he simply rejects on whim the existing semantic assignments given to the elements of his vocabulary and selects alternative ones again on whim. Humpty's speech is thus effectively unreadable to any semantic system employed by anyone other than himself. He scrambles the inputs to semantic theory into a code known only to himself. Thus communication breaks down. Applying this analogy to indexical terms, Humpty would be guilty of modifying the characters of indexicals at whim. We would not know which function from contexts to contents was encoded by his use of, e.g., "I" and hence could not calculate its content. But, as Predelli (2005, 58) explicitly states, characters are left untouched by the pre-semantic task of sentence-context pairing. Agents of answerphone messages are not using "I" to pick out anyone other than the agent of the context, they are simply selecting an improper context to pair their use of the term with.[12]

Predelli will therefore reject my interpretation of DCs as MOs.[13] There is no operator on character according to his view, only a selection of a context that we have been wrongly denied in Kaplan's theory. Dropping Kaplan's restriction to proper contexts allows us to accommodate DCs as respectable utterances made "at a distance." What, then, is to be said in favour of my view over Predelli's? I think that Predelli's view, for all its ingenuity, suffers a number of drawbacks that my account is not prone to.

Firstly, as argued in detail by Stevens (2009), Predelli's position stands or falls on the strength of the intuition that DCs really are cases where utterances

[12] An alternative source of the humpty-dumpty objection to Predelli that I have heard attacks the account on the grounds that it allows the speaker to pair a sentence with any context that they choose, hence their choices as to that pairing could, in principle, be just as private to them as Humpty's choices about meaning assignments are to his idiolect. I don't find this objection compelling—there is no obstacle to Predelli admitting that there are success conditions placed on successful communication that apply to the pairing of sentence and context just as there are for resolving lexical ambiguities. I can successfully encode a number of things by "bank," but not just *anything*. My intended meaning will only succeed if it conforms to existing conventions about English usage. Similarly for the intentions I have about the contexts I pair my sentences with.

[13] Predelli (1996) addresses the relation between monsters and DCs to some extent. Although in this paper Predelli does not consider all alleged cases of monsters, focusing solely on the famous "never put off until tomorrow what you can do today," he explicitly appeals to his intentionalist framework to explain away the apparent monstrosity of this example by maintaining that it should be understood as encoding multiple intended DCs. Discussion of his account of this particular sentence takes us beyond the scope of this paper. For other interesting discussions of cases that seem to involve one and the same sentence expressing multiple DCs as it is decoded repeatedly, see Egan (2009) and O'Madagain (2014). See Predelli (2014) for further discussion of monsters.

doi: 10.48106/dial.v74.i4.01

are made at a distance.[14] But this intuition is fragile and sensitive to varying examples. The presentation of the data makes a difference to the intuition. For example, while it is true that when I phone Amy and hear her recorded answerphone message "I am not here now, please leave a message," I understand that this expressed a prior intention on her part to communicate to anyone who hears the message the fact that she is not present at that later time, there are features one would expect an *utterance* to have which are lacking in this scenario. For example, it would be very odd of me to accuse her of lying, or even unwittingly telling an untruth, if I knew that she was in fact present at the location of her answerphone when I called her. The appropriate things to say in such a case would be something like "your *answerphone* is wrong/misleading/in need of updating, etc." not "*you* are wrong…".[15] Similar points can be made about all DCs. We intuitively recognize a gap between these devices of communication and ordinary utterances, but this gap goes intrinsically unrecognised on Predelli's intentionalist account, according to which the DC is a straightforward utterance.

Secondly, Stevens (2009) also points out that the intuition that a DC is a genuine utterance appears hard to reconcile with the equally strong intuition that an utterance is made at the time of encoding of the message. To, as Sidelle (1991, 535) puts it, engage in "arranging to make an utterance at a later time, or, if one likes, deferring an utterance," is not to engage in uttering something while making those arrangements. This is especially clear in Predelli's case.

[14] Predelli (2011) responds to the objections raised by Stevens (2009) by characterizing those objections as founded on a mistaken conception of the proper role of semantic theory. On Predelli's characterization, Stevens is denying that "the evidence put forth by true instances of 'I am not here now' [should] constrain the shape of an empirically adequate semantic account" (2009, 301). I agree that this would be a mistaken view of the role of semantic theory; however, it is a misrepresentation of the objection from Stevens, who clearly rejects the intuition that there *are* any true instances of "I am not here now" to be accounted for. Of course, Predelli is correct to note that the question of where the line should be drawn between semantic theory and pragmatic theory is a controversial one. I take the considerations in this section and the preceding one to lend compelling support to views like that forwarded in Stevens (2009) and Récanati (2010), according to which that line is decisively drawn in a way that makes pragmatic theory responsible for explaining DCs rather than semantic theory as Predelli maintains.

[15] As a referee pointed out to me, intuition seems to shift back in the other direction in a case where there is a deliberate deception. Suppose that Amy does not want to speak to me and deliberately leaves her answerphone on so that I will think she is not there. Now the intuition that she is lying has more traction. I take this point. However, I am content to use the example to illustrate that our intuitions are unstable—the intuition that agents make utterances at places and times other than where and when they are situated is malleable in a way that the intuition that agents make utterances in Kaplanian proper contexts is not.

If the context of utterance is the intended context of utterance, then I utter nothing at all when recording my answerphone message; I simply get things ready for an utterance to occur later on. Perhaps this is so, but insofar as the position is motivated by our intuitions regarding DCs, this counterintuitive consequence counts against Predelli's account. On my account, however, it can be explained easily enough. One is simply engaging in an act of pretense when recording the message, mimicking what one would say if located at the time and place of the context at which our intended audience will hear our performance. Unlike Predelli's account, this entails no claim about utterances being displaced from their proper contexts. The only utterance that takes place is that which is made when recording the message, although it is not uttered with assertoric force; it is simply the product of an act of mimicry, ready to be displayed in a different context. The underlying intuition that motivates Predelli's intentionalist account, namely the intuition that I am deliberately aiming to communicate things at contexts other than the one in which I am located, is preserved without endorsing the counterintuitive consequences of construing this as a form of utterance at a distance, by accommodating that intuition within a purely pragmatic explanation.

Technology can open up the possibility of previously outlandish uses of language but this is best explained pragmatically, not through a reconstruction of an otherwise perfectly acceptable semantic theory. Consider a recent technological advance which facilitates an unusual application of indexicals: The Rock Group, *Dio*, recently performed a series of concerts in which their deceased vocalist Ronnie James Dio was replaced on stage by a hologram. The hologram appears to be singing as it mouths along to pre-recorded vocal tracks from Ronnie. This holographic rendition of Ronnie, convincing though it may appear, is not of course really singing. The hologram is not causing any vibrations in the air, picked up by a microphone, etc. It is just a visual representation of a dead person, carefully synchronised with recordings of that person's voice. But we can exploit this pretense to the full. When performing in London, we can make our holographic Ronnie "say" things (i.e., mouth along to recorded utterances of Real Ronnie's) like "it is great to be in London tonight!". Of course, Ronnie himself is not saying anything; Ronnie, unfortunately, is dead. Suppose that on the evening of 20th December 2019, the hologram is made to "say" this sentence: "It is cold tonight in London!". Has the proposition that London is cold on the evening of 20th December 2019 been expressed? It seems reasonable to agree that it has, although it is equally obvious that Ronnie was not the agent who expressed that proposition (the

doi: 10.48106/dial.v74.i4.01

most likely agent, or agents, would be those responsible for generating and controlling the hologram). But what about a case where Holographic Ronnie "says" something using the first-person pronoun like "I am so happy to be here with you in London tonight!". Again, I think it is obvious enough that Holographic Ronnie has not said anything (not being an agent, he cannot be the agent of an utterance after all). Nor, for that matter, has Real Ronnie said that he is happy to be in London on the evening of 20th December 2019 (not being alive, he is not able to be an agent and hence not able to be an agent of an utterance). But this seems hard to square with Predelli's view, according to which the parametric settings that determine the content of an utterance are fixed by intentions. In this case there clearly is an intention, just not an intention on the part of Ronnie (Holographic or Real). But whoever produced the hologram (let us assume it is a single individual for simplicity's sake) had an intention to combine a recorded utterance of Real Ronnie's (perhaps reconstituted from several samples taken from previous utterances and hence not identical with any one previous actual utterance) with a visual representation of Real Ronnie to ensure that Holographic Ronnie "said" that he was happy to be in London on the evening of 20th December 2019.

There does not seem to be any semantic difference between what is happening here and what is happening if I attached a note authored by Jones which says "I am not here" to Smith's door to express Smith's absence. The agent of the note, on Predelli's account, is presumably Smith because it is his absence I intend to communicate. Accordingly, Predelli's account predicts that the agent of Holographic Ronnie's "utterance" is either Holographic Ronnie or, perhaps, Real Ronnie (depending on which of these two, if distinguished, the producer of the holographic performance intended). This, I think, cannot be the right thing to say in this situation. No amount of intention can make dead people agents of utterances after their death.[16] Surely what we

[16] A similar objection to Predelli is raised by Sherman (2015, 594) who notes that Predelli's intentionalist account makes apparently correct predictions about the cases where we have some choice over our use of indexicals, but struggles to explain cases where we don't. The comment is made in passing but I assume he has in mind cases like this: a recently released addition to a range of ice cream has "I am vegan!" written on it. Predelli seems to have a simple explanation of what the "I" means here—whoever wrote this intended that it be paired with a context in which the ice cream is the agent. But now consider a case where I stand next to the freezer in the supermarket shouting, "I am vegan!" to passing shoppers, while intending the sentence I emit to be paired with a context in which the ice cream is the agent. Clearly, my intention will not be fulfilled. This suggests that there is more to the successful case than just the intentions of whoever produced the communication.

have is simply a case where someone is doing an extremely sophisticated job of *pretending* that Ronnie James Dio is present and performing on stage by displaying recordings of his previous speech in an act of mimicry. This, in my view, is what we find in all cases of DCs. Ordinary utterances with Kaplanian meanings are employed to allow us to pretend to say (or pretend that others are saying) things unavailable to us when the utterances are construed literally. There is no need to modify our semantic theory to accommodate a theory of pretence.[17]

One final approach to the answerphone problem that seeks to accommodate DCs within a wholly semantic framework is suggested (though not endorsed) by Parsons (2011). The view merits brief consideration here as, again, it shares some similarities with my proposal but the differences are significant. I have argued that apparently true instances of "I am not here now" etc., are not really true. They are false utterances made by speakers who utilise a pragmatic process to facilitate their non-literal interpretation as *pretences*. Speakers are relying on context to allow them to mimic utterances at different temporal or spatial locations (or even by different agents) because that context will make those shifted contexts salient (most routinely because those shifted contexts are the ones that the hearer will be in when they decode the utterance).

It is helpful to compare my view to a radical form of what we might call "content relativism" (CR). CR is the position whereby the content of an utterance is subject to modification depending on the context in which the utterance is *assessed* (rather than the context in which it is *uttered*).[18] In Kaplanian terms,

[17] The idea that DCs can be explained as pretend utterances is also defended by Voltolini (2006) and Connolly (2017). Voltolini's strategy is to situate his explanation within a fictionalist semantics, while Connolly shares my preference for a pragmatic approach. I have a great deal of sympathy with Connolly's approach which construes DCs as produced by participants knowingly and deliberately entering into a game of "externally-oriented make believe" (2017, 616). However, while I think our approaches are in the same vein, I think the situation he describes must be supplemented by the sort of analysis I propose if it is to explain the monstrous quality of DCs. For example, I have argued that binding of indexicals (Climb *now*, work *later*), and embedded context-shifting (Amy has asked me to get her a "sick" new pair of trainers) have important similarities with DCs that require the sort of approach I am urging.

[18] CR is a more radical theory even than the controversial forms of semantic relativism (or, as it is sometimes called "truth relativism") developed most notably by Lasersohn (2017) and MacFarlane (2014). Semantic relativism holds that truth is sensitive to context of assessment; CR holds that *what is meant* by an utterance depends on the context in which that utterance is assessed. Despite the clear logical space for CR to exist within any semantic framework which admits both contexts of utterance and contexts of assessment, few have been persuaded that CR is worth exploring. A rare exception (in addition to Parsons 2011, discussed shortly) is Weatherson (2009).

doi: 10.48106/dial.v74.i4.01

the context in which an utterance is assessed is the circumstance of evaluation. Whereas Kaplan takes circumstances to be world-time pairs, CR expands the parameters to include all those parameters standardly recognized as elements of contexts of utterance (agents, times, places, etc.). Whereas circumstances of evaluation are usually appealed to in determining truth-value, CR allows them to determine content. Hence the same utterance can change its content (express a different proposition) if the context in which it is assessed changes.

Parsons (2011) considers the possibility of appealing to CR as a way of providing a semantic theory for answerphone cases. On the surface, the suggestion is promising: the utterance of "I am not here now" on the answerphone strikes us as intuitively true, despite the fact that it cannot be true if that sentence is contradictory. But, of course, CR will abandon the claim that it is contradictory, because that claim relies on the belief that the content of the sentence is tied to a proper context of utterance. CR can agree with Kaplan that all contexts of utterance are proper but take a more relaxed view on contexts of assessment, allowing these to impact the content of the sentence uttered in ways that break the tie with contexts of utterance. Hence we have a neat explanation of how an utterance of "I am not here now" can express a truth: although the sentence cannot be true when uttered, it can change its content depending on the context of its assessment so as to become true.

I think there is something right about CR, but we need to be careful about endorsing it as a semantic theory. The problem is that there is nothing systematic about the behaviour of indexicals which tells us in advance whether they are assessment-sensitive or not. Answerphone messages are assessment-sensitive, ordinary utterances of indexical sentences tend not to be. Or, to be more precise, indexical sentences uttered in certain conventionally recognised scenarios are routinely interpreted in accordance with the predictions made by CR, while most utterances do not demand such elaborate mechanisms to interpret them. Of course, we might just maintain that CR applies uniformly to all utterances but that the default interpretation is one where the context of assessment coincides with the context of utterance. Only in certain cases does the context trigger a bifurcation of context of assessment from context of utterance. I see no problem with that view, but it clearly demonstrates that the semantic theory by itself does not do sufficient explanatory work. A pragmatic account of the way in which the interaction of context of utterance and context of assessment is triggered is essential to such a story, and this is what I have attempted to provide in this paper.

One thing that suggests that a CR-based semantics alone is not sufficient to explain displaced communication is that (as we have seen many times in the discussion in this paper) our intuitions are highly unpredictable and subject to the details of the contextual situation. Parsons takes this concern to show that CR cannot explain the answerphone problem. He imagines a case where a time delay on the phone line results in someone hearing the answerphone message after the speaker has in fact returned home. With some reservation, he endorses the view that the message is still true, and is (he claims) able to shift the context of assessment away from the time of decoding to the time of *intended* decoding. Parsons himself confesses to being unsure of his intuitions in regard to this example. It seems to me to be another case like those I considered previously which just show that we do not have firm intuitions about displaced communications. But without firm intuitions to make concrete predictions about what it meant and what is true or false, the task required of a semantic theory cannot be fulfilled. The situation can be seen quite clearly by reconsidering the holographic Ronnie scenario that I posed as an objection to Predelli's approach above. It is clear that using a holographic image of Ronnie requires some rich stage-setting to work. It is only because of this stage setting that the intended content (the pretence, as I have argued) is made available. We can recognize a semantic value which is interpreted relative to the context of assessment for the utterance. This will make sense of our intuition that some of Ronnie's apparent utterances at least sound like they are true ("it is raining in Manchester tonight," for example), while others don't sound true ("I am happy to be here tonight" doesn't sound true when we know that Real Ronnie is both dead and played no conscious role in this utterance). Consider Holographic Ronnie's production of "I am Ronnie James Dio"—is this true at the context of assessment? According to CR it ought to be possible that the agent really will be Real Ronnie. And Real Ronnie really is Ronnie James Dio. So the utterance should be true. But I don't have the intuition that this utterance is true—or, rather, I'm not sure that I have *any* intuition about this sort of case. Intuitions are just not stable in cases like these. And unstable intuitions are not suitable foundations for a semantic theory.

6 Objections and Replies

In this final section, I will consider some objections to the view that I have presented above, and offer some replies, which will hopefully help to clarify my position.

The first objection I want to consider concerns my definition of DCs. A DC is a communication that occurs at a different context to that in which it is encoded. It is tempting to assume (as seems to be the case for each of the examples considered so far) that DCs are *always* evaluated with respect to the context in which they are decoded (hence, on my view, the monstrosity present in the pragmatic operation facilitating DCs) But what about cases like we see in the following pair (both, imagine, recorded for a posthumously broadcasted will):

(r) Today, I met with my lawyer before recording this will.
(s) Today, you all received a call from my lawyer informing you that you have inherited a large sum.

It seems that both (r) and (s) are clear cases of DCs as commonly discussed in the literature, yet only (s) seems to communicate information that is evaluated with respect to the context in which it is decoded. Far from being monstrous, (r) seems to communicate information about the context of encoding. But is this not a DC?

I do not think that (r) is a DC. While (r) is being used to communicate information at a context subsequent to that in which it is encoded, the information is about the context of encoding. The indexicals “today,” “I,” “my,” and “this” all contribute contents drawn from the context in which the message is recorded. Furthermore, I am sceptical that a construction like (r) could be developed in such a way as to be coherently understood as communicating information about the context of decoding. For example, continuing (r) in the following way, sounds infelicitous to my ear:

(r*) Today, I met with my lawyer before recording this will that you are now listening to.

If we understand this as an attempt to shift the temporal parameter of the utterance from that indicated by “today” to that indicated by “now,” mid-sentence, I think the sentence can only be made sense of if we read an implicit open quotation as present on the “now.” Only in such cases, I suggest, do we have a candidate for a DC. Simply presenting a recording of a message is not sufficient to produce a DC. Only when that message is naturally interpreted as communicating information *about* the context in which it is decoded, rather than encoded does it count as a DC. If I uncover a forgotten recording from my 10th birthday in which I say "I am 10 today," I do not stumble on a DC. But if I uncover a recording of my 10-year-old self, saying "when you hear this, you will suddenly remember recording it when you were 10," I do. It seems to me that (r) is akin to the former, not the latter.

The second objection arises when we consider a very large class of cases of potential DCs that I have said little about above, involving the production of *signs* containing indexicals. Consider a sign positioned in a hospital waiting room that consists of an inscription "please wait here." This sign exhibits typical features of a DC as "here" will be naturally interpreted as referring to the location of installation, not of inscription. Tokens of the type of this sign are mass-produced in a factory. Some individual factory worker produced this particular token sign. But, surely, the producer of the sign in this case is not the agent of any instruction. The factory worker is simply a component part in the production of a communication that intuitively seems to occur at the time of decoding. This potentially casts doubt on my claim that DCs are the result of MOs operating on an utterance or inscription evaluated with respect to a proper Kaplanian context.[19]

I agree that it is implausible to construe the factory worker as the agent of the instruction inscribed on the sign. I am, however, unconvinced that *any* instruction as such is made in the factory. We should not be misled by the fact that human agents can be involved in the production of an artefact that carries information into inferring that they are the agents of whatever information is thereby transmitted. In this instance, the factory worker is no more the agent of a communicated content, than Stephen Hawkins's voice-synthesising computer is the agent of his utterances when he relies on it to communicate his thoughts. The factory worker is producing another agent's message in accordance with their instructions. Who, then, is the agent who desires to communicate the information? The agent here is the hospital (or relevant

[19] Examples such as these motivate both O'Madagain (2014) and Briciu (2018) to distinguish between tokens and proper utterances. This distinction allows for the possibility of utterances at a distance by holding that genuine utterance requires the presence of illocutionary force, whereas the mere production of a token does not. I am inclined to agree that, in the example above, our factory worker is engaged in the production of a token, not an utterance but, as I now argue, I do not think that this means we must recognize the context of decoding as the context of utterance.

hospital authority). We can avoid complicated metaphysical questions about how organisations might be agents by assuming an individual consultant, Ms Smith, is the relevant authority. Ms Smith wants to ensure that patients arriving in reception wait in an orderly fashion in the waiting room. One way that she could do this would be to write a sign in her own hand saying “please wait here” ready to be displayed at the waiting room, or utter the sentence “please wait here” into a recording device to be on looped playback in the reception. But, due to the frequent reoccurrence of episodes when consultants need to instruct patients to wait in a specific location, it is of course more practical for signs to be mass-produced rather than produced by flimsy hand-written notes. Hence she orders a batch of ready-made signs designed to meet this common need among consultants. Nonetheless, Ms Smith remains the agent of the instruction. She has simply exploited a labour-saving device that ensures that one factory worker produces signs for the large number of agents who want to issue this instruction. Once she is in possession of the sign, she can exploit the convention that signs routinely signal information about their spatial location to engage in pretence of the sort her hand-written note would exploit.[20] The difference in the method of production of her message does not alter the fact that she is the agent of the utterance and its displacement is the result of a metalinguistic pretence, not a deferral of her utterance.

Another objection responds directly to my analysis of open quotation as an MO. An obvious feature of open quotation is that, even if the quotation operation is not explicit, it should be easily recoverable. Consider this exchange from the movie *The Empire Strikes Back*. Lando Calrissian has double-crossed Han Solo and his friends, betraying them to Darth Vader and the evil Empire. However, he strikes a deal with Vader to preserve the freedom of Solo’s friends. Informing Solo of the deal, he says “I”ve done all I can. I’m sorry I can’t do more, but I’ve got my own problems.” Solo sarcastically replies: “Yeah. You’re a real hero.” It is obvious enough how we might appeal to an open quotation analysis of this ironic utterance. Solo is not expressing his own admiration for

[20] Not all signs are obviously about the location in which they are placed, or object they are attached to, of course. An object may well be emblazoned with the sign “visit [such and such website] to see full product range,” or a pair of running shoes may come in a box marked “consult medical professional before beginning any new program of exercise.” Such signs, while clearly connected to some salient object are not about that object. But there are clearly a multitude of cases where the convention does hold: “twist clockwise” on a food jar lid, “made in England” on a guitar amplifier, “serve chilled” on a beer bottle, “4m high” on a road bridge, etc., all refer to the object they are attached to. “No smoking” in a public building, “Slow Down” on a road sign, “Wear a face covering” outside a shop, etc., all refer to the location in which they are placed.

Calrissian (he has in fact just punched Calrissian in the face, unequivocally expressing his real attitude). He does not mean that Calrissian is a hero, rather he is mockingly echoing the use of this term of praise to display his own distance from such a perspective. One thing that obviously stands in favour of the open quotation analysis is that the recovery of the operation as an explicit one is natural. One might very well report Solo's speech as "Yeah, You're a real 'hero'." Many of the commonly cited cases of DCs in the literature on indexicality, however, do not seem to be so neatly reconfigured with explicit quotation marks. Consider Predelli's note from section 1:

> I am not at home now. If you hurry, you'll catch the evening flight to Los Cabos. Meet me in six hours at the Hotel Cabo Real.

It would not be natural to add quotation marks to the shifted indexicals in the note (I add the "#" to indicate the marked quality of this):

> #I am not at home "now." If you hurry, you'll catch the evening flight to Los Cabos. Meet me "in six hours" at the Hotel Cabo Real.

The note, if anything, becomes quite confusing once the quotation marks are made explicit. Why is this, and how can it be the case if a DC is really generated by open quotation, in the same way as Han's response to Lando?

A key difference is apparent in these two contrasting cases that explains why quotation is not recoverable in the second case. The first case involves the shifting of a sub-sentential element within a context that remains non-shifted. The second case involves the shifting of the entire sentence for its interpretation. But open quotation is linguistically employed for the first kind of case only. Recall that open quotation is appealed to on my account as a way of making explicit a form of mimicry. In a case where a single expression, or string of expressions, contained in a wider linguistic frame are employed in this mimicking role while the wider frame is not, open quotation serves to explicitly indicate this role. When it is an entire sentence or other self-contained linguistic item, this device serves no purpose. Mimicry shifts the context to create a DC. Only when mimicry is embedded within a non-shifted context is explicit quotation required to indicate this. This is why, for example, it makes no sense to add quotation marks to this message:

(t) "I'll be back"

doi: 10.48106/dial.v74.i4.01

But they are clearly useful in:

(t*) As Arnold Schwarzenegger would say, "I'll be back."

In Predelli's note, we cannot capture the mimicry that I take it to employ by only quoting "now" and "in six hours," because they have not been shifted in relation to the remainder of the note. It is the whole note that mimics what the note writer intends to communicate by their pretence. My claim is that the same sort of pretence lies behind both cases, but only when the mimicry is embedded within a wider frame is the quotation device demanded to make this explicit.

The final objection I want to consider is a methodological one that, I think, goes to the heart of the different approaches to this problem taken by Predelli and those who, like myself, have urged a pragmatic explanation of DCs. The intuition that supports the pragmatic explanation of DCs is that distinctive features of the concrete episodes of language use that generate DCs are responsible for them. These features point to ways in which DCs are heavily reliant on a rich background of human behaviour that goes beyond the remit of semantic theory to explain. Like other aspects of communication that are accepted as requiring an explanation from pragmatics such as conversational implicatures, we need to look beyond the literal meanings of the expressions employed to understand what is happening in these cases. But, insists Predelli (2005, 2011), this approach both misunderstands and underestimates the place of semantics. It fails, in his view, to respect Kaplan's advice that we base our semantic theorizing on "the verities of meanings," not "the vagaries of actions" (1989b, 585). Indeed, Kaplan's own insistence on admitting only proper contexts is a failure to follow his own advice, according to Predelli (2005, 60–62). The view that Kaplan's restriction on contexts strays from is summarised elsewhere by Predelli (2011, 301) like this: "[S]emantics is concerned with the evaluation of sentences with respect to contexts, not with concrete episodes of language use—it is concerned with 'utterances' only in Kaplan's technical sense of the term as sentence-context pairs". To illustrate the significance of drawing this methodological line, Predelli gives the example of a tautology that is sufficiently long that no mortal human will ever utter it. As a concrete episode of language use, we do not have an utterance. But clearly it is unproblematic to evaluate the sentence as true with respect to any context of "utterance," in the more careful Kaplanian sense of a sentence-context pair (see 1989a, 522–523). Failing to respect this distinction, and being misled by the peculiarities of how we use language to perform speech acts in particular situations, Predelli maintains, inevitably leads us to the wrong conclusions.

While I agree with Predelli that we ought to recognise the distinction he makes, I do not agree that there is a methodological decision to be taken here that will remain neutral with regard to our intuitions concerning concrete episodes of language use. Keeping the discussion focused purely on the issue at hand, one can of course construct a formal semantic theory that is more generous in the contexts it evaluates sentences with respect to than the proper contexts endorsed by Kaplan. One is limited only by mathematical constraints in this regard. But eventually one has to make a decision about which of those mathematical possibilities correspond to our *actual* use of natural language expressions, if the mathematical structure in our formal semantics is going to be empirically adequate as a model of the semantic profile of an actual expression or set of expressions in a natural language. Kaplan's decision to restrict the range of contexts we should be interested in to proper ones is, I take it, based on this desideratum. After noting that an unconstrained range of contexts will provide contexts with respect to which "I am here now" is false, he insists that only the proper ones should be admitted if we are to arrive at an empirically adequate analysis of the indexical expressions contained in this sentence. To repeat the quotation I gave at the beginning of this paper: "[I]mproper indices are like impossible worlds; no such contexts could exist and thus there is no *interest* in evaluating the extensions of expressions with respect to them" (Kaplan 1989a, 509, emphasis added). As the quotation shows, while Predelli is quite correct to point out that a formal semantic theory should be founded on an abstract pairing of expressions with mathematical objects within a formal structure, we will have to make choices about which pairings are of interest to our concerns as natural language semanticists, and these choices will surely be based on our intuitions about the way the expressions behave in the mouths, pens, and thoughts of ordinary speakers. My own intuition, following Kaplan's, is that the restriction of contexts to the set of proper contexts best captures the semantic behaviour of indexicals in English, once we recognise the input of pragmatic processes on apparent deviations from this restriction. Predelli's intuition is to take the deviations to illustrate that Kaplan's restriction is empirically inadequate. I do not think that either of us is basing our choice about which contexts our semantic theories should recognise on issues independent of intuitions about concrete episodes of language use, nor do I think that we should.

7 Conclusion

My argument in this paper has been the following. Firstly, I have argued that DCs are best understood as being generated by MOs. I have then argued that, understood as MOs, they are in turn best understood as the result of pragmatically triggered metalinguistic context-shifting operations. I have then given a detailed explanation of this proposed mechanism. If this is correct, then DCs are MOs but are not Monsters, for, while DCs are certainly monstrous, their monstrosity is not generated by any lexicalised semantic operator of English. Furthermore, the argument presented here is also intended to vindicate Kaplan's insistence that the only proper contexts relevant to the semantic evaluation of English indexicals are those which situate the agent at the time and place of her utterance. DCs are not "utterances at a distance" which result from making utterances in improper contexts; they are ordinary utterances made in the course of a deliberate pretence that they are something more.[*]

Graham Stevens
0000-0003-3832-6391
University of Manchester
Graham.P.Stevens@manchester.ac.uk

References

Åkerman, Jonas. 2017. "Indexicals and Reference-Shifting: Towards a Pragmatic Approach." *Philosophy and Phenomenological Research* 95(1): 117–152, doi:10.1111/phpr.12216.

Bianchi, Claudia. 2014. "Slurs and Appropriation: An Echoic Account." *The Journal of Pragmatics* 66: 35–44, doi:10.1016/j.pragma.2014.02.009.

Briciu, Adrian. 2018. "Indexicals in Remote Utterances." *Philosophia* 46(1): 39–55, doi:10.1007/s11406-017-9909-x.

Cappelen, Herman and LePore, Ernest. 1997. "The Varieties of Quotation." *Mind* 106(424): 429–450, doi:10.1093/mind/106.423.429.

Chuck D and Jah, Yusuf. 1998. *Fight the Power: Rap, Race, and Reality*. New York: Dell Publishing.

Cohen, Jonathan. 2013. "Indexicality and the Puzzle of the Answering Machine." *The Journal of Philosophy* 110(1): 5–32, doi:10.5840/jphil2013110143.

Connolly, Niall. 2017. "I Am Here Now, But I Won't Be Here When You Get This Message." *Dialectica* 71(4): 603–622, doi:10.1111/1746-8361.12208.

[*] THANKS

CORAZZA, Eros, FISH, William and GORVETT, Jonathan. 2002. “Who is ‘I’?” *Philosophical Studies* 107(1): 1–21, doi:10.1023/A:1013111419036.

DAVIDSON, Donald. 1979. “Quotation.” *Theory and Decision* 11(1): 27–40. Reprinted in Davidson (1984, 479–488), doi:10.1007/BF00126690.

—. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

EGAN, Andy. 2009. “Billboards, Bombs and Shotgun Weddings.” *Synthese* 166(2): 251–279, doi:10.1007/s11229-007-9284-4.

KAPLAN, David. 1989a. “Demonstratives.” in *Themes from Kaplan*, edited by Joseph ALMOG, John R. PERRY, and Howard K. WETTSTEIN, pp. 481–563. Oxford: Oxford University Press. Widely circulated from 1977 on.

—. 1989b. “Afterthoughts.” in *Themes from Kaplan*, edited by Joseph ALMOG, John R. PERRY, and Howard K. WETTSTEIN, pp. 565–614. Oxford: Oxford University Press.

KING, Jeffrey C. 2001. *Complex Demonstratives: A Quantificational Account*. Cambridge, Massachusetts: The MIT Press.

LASERSOHN, Peter. 2017. *Subjectivity and Perspective in Truth-Theoretic Semantics*. Oxford Studies in Semantics and Pragmatics n. 8. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199573677.001.0001.

MACFARLANE, John. 2014. *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199682751.001.0001.

O’MADAGAIN, Cathal. 2014. “Indexicals and the Metaphysics of Semantic Tokens: When Shapes and Sounds become Utterances.” *Thought* 3(1): 71–79, doi:10.1002/tht3.114.

PARSONS, Josh. 2011. “Assessment-Contextual Indexicals.” *Australasian Journal of Philosophy* 89(1): 1–17, doi:10.1080/00048400903493530.

PREDELLI, Stefano. 1996. “Never Put Off Until Tomorrow What You Can Do Today.” *Analysis* 56(2): 85–91, doi:10.1093/analys/56.2.85.

—. 2005. *Contexts. Meaning, Truth, and the Use of Language*. Oxford: Oxford University Press, doi:10.1093/0199281734.001.0001.

—. 2011. “I Am Still Not Here Now.” *Erkenntnis* 74(3): 289–303, doi:10.1007/s10670-010-9224-4.

—. 2014. “Kaplan’s Three Monsters.” *Analysis* 74(3): 389–393, doi:10.1093/analys/anu059.

RÉCANATI, François. 2010. *Truth-Conditional Pragmatics*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199226993.001.0001.

SCHLENKER, Philippe. 2003. “A Plea for Monsters.” *Linguistics and Philosophy* 26(1): 29–120, doi:10.1023/A:1022225203544.

SCOTT, Michael and STEVENS, Graham. 2019. “An Indexical Theory of Racial Pejoratives.” *Analytic Philosophy* 60(4): 385–404, doi:10.1111/phib.12156.

SHERMAN, Brett. 2015. “Constructing Contexts.” *Ergo* 2(23): 581–605, doi:10.3998/ergo.12405314.0002.023.

Sidelle, Alan. 1991. “The Answering Machine Paradox.” *Canadian Journal of Philosophy* 21(4): 525–539, doi:10.1080/00455091.1991.10717260.

Stevens, Graham. 2009. “Utterance at a Distance.” *Philosophical Studies* 143(2): 213–221, doi:10.1007/s11098-007-9199-4.

Stevens, Graham and Duckett, Nathan. 2019. “Expressive Content and Speaker-Dependence.” *Linguistic and Philosophical Investigations* 18: 97–112, doi:10.22381/LPI1820195.

Voltolini, Alberto. 2006. “Fiction as a Base of Interpretation Contexts.” *Synthese* 153(1): 23–47, doi:10.1007/s11229-006-0001-5.

Weatherson, Brian. 2009. “Conditionals and Indexical Relativism.” *Synthese* 166(2): 333–357, doi:10.1007/s11229-007-9283-5.

Wilson, Deirdre. 2006. “The Pragmatics of Verbal Irony: Echo or Pretence?” *Lingua* 116(10): 1722–1743, doi:10.1016/j.lingua.2006.05.001.

Constitutivism about Instrumental Desire and Introspective Belief

Ryan Cox

This essay is about two familiar theses in the philosophy of mind: constitutivism about instrumental desires, and constitutivism about introspective beliefs, and the arguments for and against them. Constitutivism about instrumental desire is the thesis that instrumental desires are at least partly constituted by the desires and means-end beliefs which explain them, and is a thesis which has been championed most prominently by Michael Smith. Constitutivism about introspective belief is the thesis that introspective beliefs are at least partly constituted by the mental states they are about, and is a thesis which has been championed most prominently by Sydney Shoemaker. Despite their similarities, the fortunes of these two theses could not be more opposed: constitutivism about instrumental desire is widely accepted, and constitutivism about introspective belief is widely rejected. Yet, the arguments for both theses are roughly analogous. This essay explores these arguments. I argue that the argument which is widely taken to be the best argument for constitutivism about instrumental desires—what I call the argument from necessitation—does not provide the support for the thesis it is widely taken to provide, and that it fails for much the same reasons that it fails to provide support for constitutivism about introspective belief. Furthermore, I argue that the best argument for constitutivism about instrumental desires—what I will call the argument from cognitive dynamics—is also a good argument, if not equally good, for constitutivism about introspective belief (at least when the thesis is suitably qualified).

This essay is about two familiar theses in the philosophy of mind: *constitutivism about instrumental desire*, and *constitutivism about introspective belief*, and the arguments for and against them. Constitutivism about instrumental desire is the thesis that instrumental desires are at least partly constituted by the desires and means-end beliefs which explain them and is a thesis which has been championed most prominently by Michael Smith (2004). Constitutivism about introspective belief is the thesis that introspective beliefs are at least partly constituted by the mental states they are about and is a thesis which has been championed most prominently by Sydney Shoemaker (1996, 2012). Despite their similarities, the fortunes of these two theses could not be more opposed: constitutivism about instrumental desire is widely accepted, and constitutivism about introspective belief is widely rejected. Yet, the arguments for both theses are roughly analogous. So if one thesis is to be accepted while the other is rejected there must be good reasons for rejecting the arguments for one thesis but not the other.

In this essay, I argue that the *best* argument for constitutivism about instrumental desire—what I will call the *argument from cognitive dynamics*—is also a *good* argument, if not an *equally good* argument, for constitutivism about introspective belief (at least when the thesis is suitably qualified). So, at least with respect to this argument, there are no good reasons for accepting one thesis while rejecting the other. At the same time, however, I argue that the argument which is widely taken to be the best argument for constitutivism about instrumental desire—what I will call the *argument from necessitation*—does not provide the support for the thesis it is widely taken to provide, and that it fails for much the same reasons that it fails to provide support for constitutivism about introspective belief. So, with respect to this argument, there are no good reasons for accepting one thesis while rejecting the other, because this argument does not give us good reasons for accepting either thesis.

These conclusions suggest that the fortunes of constitutivism about instrumental desire and constitutivism about introspective belief are more closely tied together than is often appreciated. Indeed, I hope to bring to bear on the topic of instrumental desire an important lesson which has been learnt in the philosophy of introspection. For philosophers of introspection have shown that the argument from necessitation for introspective belief is unsuccessful. This partly explains why constitutivism about introspective belief is not widely accepted in the way constitutivism about instrumental desire is. Yet an analogous lesson for the argument from necessitation for constitutivism about instrumental desire has not yet been absorbed by those working on the philosophy of instrumental desire. There is a certain irony here, since Shoemaker explicitly draws analogies with the case of instrumental desire—particularly the role played by means-end beliefs and non-instrumental desires in rationalising and explaining further desires—in developing his arguments for constitutivism about introspective belief (1996). Rejecting this argument for constitutivism about instrumental desire puts this thesis on a less firm footing, and brings its fortunes more closely into line with those of constitutivism about introspective belief. As I will argue, the fortunes of both theses rest on the prospects of the argument from cognitive dynamics. I hope to show that, even if the argument from cognitive dynamics is taken to offer adequate support for constitutivism about instrumental desire while ultimately not offering adequate support for constitutivism about introspective belief, the fortunes of these two theses are tied closer together than is often appreciated.

The essay is structured as follows. Section 1 discusses some preliminary issues concerning how constitutivism about instrumental desire and constitutivism about introspective belief are to be understood. Section 2 considers the argument from necessitation for constitutivism about instrumental desire and the argument from necessitation for constitutivism about introspective belief respectively. I argue that both arguments fail. Section 3 considers the argument from cognitive dynamics for constitutivism about instrumental desire and the argument from cognitive dynamics for constitutivism about introspective belief respectively. I argue that the arguments provide equally good reasons for accepting both theses.

1 Locating the Topic

At the outset, we are going to need a way of understanding instrumental desires and introspective beliefs which does not prejudice the case either for or against the respective constitutivist theses. It might be thought, after all, that there isn't much to say about constitutivism about instrumental desire, since instrumental desires just are, by definition, according to some, those desires which are at least partly constituted by the desires and means-end beliefs which explain them. While the term "instrumental desire" is sometimes used this way, there is a more neutral way of understanding instrumental desires which does not prejudice the issue in this way. We can simply say that instrumental desires just are, by definition, those desires which are *rationally explained* by other desires and means-end beliefs, where we leave open the question of whether the former is partly constituted by the latter. This neutral understanding of instrumental desire is widespread in the literature (Marks 1986, 9; Davis 1986, 69; Schroeder 2004, 5; McDaniel and Bradley 2008, 286; Arpaly and Schroeder 2014, 6). It is arguably this understanding which Hume has in mind when he writes:

doi: 10.48106/dial.v74.i4.02

> Ask a man *why he uses exercise*; he will answer *because he desires to keep his health*. If you then enquire, *why he desires to keep his health*, he will readily reply, *because his sickness is painful*. If you push your enquiries farther, and desire a reason *why he hates pain*, it is impossible he can ever give any. This is an ultimate end, and is never referred to any other object. (Hume EPM, Appendix I)[1]

In this way, Hume distinguishes between what we would call the instrumental desire to exercise, rationally explained by a desire to keep one's health, and the non-instrumental desire to avoid pain, not explained by any further desire. When Smith writes "I will call the desires which are explained by [non-instrumental] desires and means-end beliefs 'instrumental' desires" (2004, 95) he is clearly stating that he means to use "instrumental desires" in this neutral way.[2]

On this way of understanding instrumental desires, it is clearly a substantive question whether instrumental desires are partly constituted by means-end beliefs and other desires. Of course, many theorists accept this substantive thesis about instrumental desires. Here is Smith's assertion of his commitment to constitutivism about instrumental desire:

> Instrumental desires are not distinct from the non-instrumental desires and means-end beliefs that explain them, but are rather just the complex state of having such non-instrumental desires and means-end beliefs standing in a suitable relation. (2004, 96)

According to Smith, then, instrumental desires are (at least partly, if not wholly) constituted by non-instrumental desires and means-end beliefs standing in suitable relations. As he puts it: "Instrumental desires are thus better thought of as being nothing over and above the non-instrumental desires and means-end beliefs that explain them" (2004, 96).[3] Here Smith appears to commit himself to a strong form of constitutivism about instrumental desire, one which holds that instrumental desires are "nothing over and above" or are *wholly constituted* by the non-instrumental desires and means-end beliefs that

[1] Quoted in Smith (2004, 94).

[2] See also "[I]nstrumental desires are those that can be explained by non-instrumental desires and means-end beliefs" (Smith 2004, 96–97).

[3] The qualification that the non-instrumental desire and the means-end belief must stand in a suitable relation is missing here but we can assume that it is intended.

explain them standing in a suitable relation. Constitutivism about instrumental desire, so understood, is a widely accepted thesis in the philosophy of mind. However, as we have just seen, it is a substantive thesis about instrumental desires, and we will be concerned with the arguments for it in the next two sections.

Introspective beliefs raise their own unique problems for understanding. While there is no temptation in the case of introspective beliefs to hold that they are, by definition, beliefs which are partly constituted by the mental states they are about, it is nonetheless difficult to say what an introspective belief is. It is not enough to say that an introspective belief is just a belief about one's own mental states. For, it is relatively uncontroversial that at least some of our beliefs about our own mental states are arrived at on the basis of inference. And these beliefs are very plausibly thought of as being constitutively distinct from the mental states they are about. They are no more constituted by the mental states they are about than our beliefs about the mental states of others are constituted by the mental states they are about. Rather, introspective beliefs must be understood as beliefs about our own mental states which are arrived at by some special means and are not based on evidence or observation in the way that our beliefs about the mental states of others are.[4] While it is a matter of controversy how introspective beliefs are to be understood, I will assume that some such distinction can be drawn among beliefs about our own mental states and that only some of these beliefs will count as introspective beliefs. Constitutivism about introspective belief holds that *these* beliefs are at least partly constituted by the mental states they are about. Here is Shoemaker's assertion of his commitment to constitutivism about introspective belief:

> What I am inclined to say is that second-order belief, and the knowledge it typically embodies, is supervenient on first order beliefs and desires—or rather, it is supervenient on these plus a certain degree of rationality, intelligence, and conceptual capacity. By this I mean that one has the former *in* having the latter—that having the former is nothing over and above having the latter. (1996, 34)

[4] For an influential discussion of these features of introspective beliefs see Moran (2001). For scepticism about the existence of introspective beliefs so understood see Cassam (2014).

doi: 10.48106/dial.v74.i4.02

According to Shoemaker, then, introspective beliefs are at least partly (if not wholly) constituted by the mental states they are about together with "a certain degree of rationality, intelligence, and conceptual capacity"—having the former is "nothing over and above" having the latter. This passage anticipates the argument to be examined in the next section, as Shoemaker moves here from a claim about supervenience or necessitation to a claim about constitution. So we have seen that constitutivism about introspective belief is a substantive thesis about introspective belief. It will be the aim of the next sections of the essay to evaluate arguments for both forms of constitutivism.

Before turning to those arguments a final qualification is in order. In the passage from Shoemaker just quoted, Shoemaker is concerned with a kind of constitutivism about introspective belief which concerns introspective beliefs about *attitudes* in particular. For the most part, in what follows, I will be concerned with forms of constitutivism about introspective belief which are restricted in this way, holding that introspective beliefs about our *attitudes* are partly constituted by the attitudes they are about. Introspective beliefs about phenomenally conscious states raise further issues that I will not be able to address here, and while constitutivism about these introspective beliefs may be defensible, different arguments may be required.[5] There is also an interesting question about how constitutivism might be extended to states which are plausibly thought of as having both cognitive and non-cognitive components, like emotions. While I think that constitutivism can be defended for a wide range of mental states, I will largely set aside such an exploration here, and will focus on the particular case of introspective beliefs about our own attitudes. I will also set aside the difficult question of whether we can or should expect a uniform account of introspection and introspective beliefs which applies to all mental states.[6] Finally, related to these questions is the question of which theories of introspection—theories of the means by which we arrive at introspective beliefs—are compatible with constitutivism about introspective belief and those which are not. Some theories of introspection have implications for the relation between introspective beliefs and the mental states they are about. The self-scanning theory of David Armstrong (1968) has the implication that introspective beliefs are constitutively distinct from the states they are about since the former are caused by the latter. Other theories are neutral about the relation and while they may provide causal

[5] See Chalmers (2010) for a discussion of phenomenal beliefs.
[6] See Boyle (2009) and Byrne (2011) for discussion.

explanations of the means by which we arrive at introspective beliefs—by means of answering deliberative questions (Moran 2001), or by means of an ascent-routine (Gordon 1986)—they remain compatible with the possibility that introspective beliefs are partly constituted by the attitudes they are about. I will also set aside this question, as the arguments we will consider for constitutivism about introspective belief do not presuppose any particular theory of introspection.

2 The Arguments from Necessitation

In this section I will formulate and evaluate the argument from necessitation for constitutivism about instrumental desire and the argument from necessitation for constitutivism about introspective belief. The arguments belong to a family of arguments, arguments from necessitation or supervenience, which are familiar enough across many areas of philosophy. The crucial step in such arguments is a move from a claim about metaphysical necessitation or supervenience, to a claim about constitution. I will first formulate and motivate each argument and then turn to evaluation.

2.1 *For Constitutivism About Instrumental Desire*

For simplicity, we can formulate and motivate the argument from necessitation for constitutivism about instrumental desire by focussing on an arbitrary example which we can take to reveal something general about instrumental desires. Suppose, then, that Jane desires to exercise because she desires to keep her health, and believes that exercising is a means to keeping her health. It follows from our understanding of instrumental desires, and the assumption that this is the "because" of rational explanation, that Jane instrumentally desires to exercise. Why think that Jane's desiring to exercise is at least partly constituted by her desiring to keep her health and her believing that exercising is a means to keeping her health? The argument from necessitation proceeds in two steps.

The first step establishes that the relation between Jane's desiring to keep her health, her believing that exercising is a means to keeping her health, *her being fully rational*, and her desiring to exercise is not a merely contingent relation, but is, in some sense, necessary. We can bring this out by reflecting on a claim about necessity like the following:

> N_1. Necessarily, if Jane desires to keep her health, believes that exercising is a means to keeping her health, and is fully rational, then she desires to exercise.

I will have more to say about the relevant understanding of "is fully rational" in this claim below, and will connect this to Smith's claims about means-end beliefs and non-instrumental desires standing in suitable relations. But for now, we can simply observe that on a natural understanding of "is fully rational" such a claim is intuitively plausible. To bring this out we might notice that while it is certainly possible for Jane to desire to keep her health, to believe that exercising is a means to keeping her health, while not desiring to exercise to any degree—after all, she might be less than fully rational—it is not possible for her to desire to keep her health, to believe that exercising is a means to keeping her health, and not desire to exercise to any degree *if she is fully rational.*

The second step establishes the best explanation of the necessary connection is that Jane's desiring to exercise is at least partly, if not wholly, constituted by her desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational. While it might be tempting to move directly from the claim about necessitation to this conclusion, it is important to see that there are alternative explanations of the necessary connection which need to be considered. The explanations of the necessary connection we need to consider in this case are these:

> H_1. Jane's desiring to exercise is wholly constituted by her desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational.

> H_2. Jane's desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational is partly constituted by her desiring to exercise.

> H_3. There is something that Jane's desiring to exercise is wholly constituted by and which Jane's desiring to keep her health, her believing that exercising is a means to keeping her health and her being fully rational is partly constituted by.

These explanations correspond to the familiar options for explaining necessary connections of this form. If there is a necessary connection between something's being an *F* and its being both a *G* and an *H*, then we might explain this necessary connection by holding that its being an *F* is wholly constituted by its being a *G* and an *H*, or by holding that its being a *G* and an *H* is partly constituted by its being an *F*, or by holding that there is something else, such that its being a *G* and an *H* is partly constituted by and its being an *F* is wholly constituted by.

Of these explanations, H_1 certainly looks to be the best. There is no candidate for the kind of third-factor required by H_3, and, at least initially, it is hard to see why Jane's desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational, would be partly constituted by her desiring to exercise: none of these conditions seem to be independently partly constituted by her desiring to exercise, and it is hard to see how, jointly, they could be partly constituted by her desiring to exercise. So we may tentatively conclude that H_1 provides the best explanation of the necessary connection.

These two steps, then, provide motivation for the two premises of the argument from necessitation for constitutivism about instrumental desire. We can think of the argument as proceeding as follows:

P_1 There is a necessary connection between (i) Jane's desiring to exercise and (ii) her desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational.

P_2 The best explanation of this necessary connection is that Jane's desiring to exercise is wholly constituted by her desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational.

C_1 Jane's desiring to exercise is wholly constituted by her desiring to keep her health, her believing that exercising is a means to keeping her health, and her being fully rational.

Since the example we have focused on here was entirely arbitrary, the same reasoning can be followed in arguing for constitutivism about instrumental desire as a general thesis. At least initially, then, the argument from necessitation provides a good case for constitutivism about instrumental desire. I will now explicate an analogous argument for constitutivism about introspective belief before turning to objections.

2.2 *For Constitutivism About Introspective Belief*

For simplicity, we can formulate and motivate the argument from necessitation for constitutivism about introspective belief by focusing on an arbitrary example which we can take to reveal something general about introspective beliefs. Suppose that Jane introspectively believes that she believes it is about to rain. Why think that Jane's believing that she believes it is about to rain is not constitutively distinct from her believing that it is about to rain? The argument proceeds in two steps.

The first step establishes that the relation between Jane's believing that she believes that it is about to rain and her believing that it is about to rain, her having some interest in the question of whether she believes that it is about to rain, and her being fully rational, is not a merely contingent relation, but is, in some sense, necessary. We can bring this out by reflecting on a claim about necessity like the following:

> N_2. Necessarily, if Jane believes that it is about to rain, understands and has some interest in the question of whether she believes that it is about to rain, and is fully rational, then she believes that she believes that it is about to rain.

A few clarifications are in order here. Recall Shoemaker's claim that introspective beliefs are supervenient on "a certain degree of rationality, intelligence, and conceptual capacity." This suggests a claim like the following: necessarily, if Jane believes that it is about to rain, has a certain degree of rationality, intelligence, and conceptual capacity, she believes that it is about to rain. Shoemaker adds these claims about intelligence and conceptual capacity here in order to avoid problems stemming from small children and animals who may have the relevant attitudes, have a certain degree of rationality, and yet not even be able to understand the question of whether they have the attitudes in question. I have captured this element of Shoemaker's view with the claim about understanding in N_2. However, I have added the further claim that Jane must *have some interest in the question of whether she believes that it is about to rain*. Arguably, Shoemaker's conditions are too weak. It seems to be possible for Jane to believe that it is about to rain, for her to understand the question of whether she believes that it is about to rain, for her to be fully rational, and yet for her not to believe that she believes that it is about to rain *if she* *has no interest in the question of whether she believes that it is about to rain.*[7] Quite generally, it seems that we will not believe that we have some attitude or another when we have no interest in the question of whether we have that attitude. With these clarifications in order, we can see that this claim is intuitively plausible on a natural understanding of "is fully rational." To bring this out we might notice that while it is certainly possible for Jane to believe that it is about to rain, to have some interest in the question of whether she believes that it is about to rain, while not believing that she believes that it is about to rain—after all, she might be less than fully rational—it is not possible for her to believe that it is about to rain, to have some interest in the question of whether she believes that it is about to rain, and to not believe that she believes that it is about to rain *if she is fully rational.*

The second step in the argument establishes that the best explanation of this necessary connection is that Jane's believing that she believes that it is about to rain is at least partly, if not wholly, constituted by her believing that it is about to rain, her having some interest in the question of whether she believes that it is about to rain, and her being fully rational. Again, we must consider the alternative explanations. The explanations of the necessary connections we need to consider in this case are these:

> H_1. Jane's believing that she believes that it is about to rain is wholly constituted by her believing that it is about to rain, her understanding and taking an interest in the question of whether she believes that it is about to rain, and her being fully rational.
>
> H_2. Jane's believing that it is about to rain, her understanding and taking an interest in the question of whether she believes that it is about to rain, and her being fully rational is partly constituted by her believing that she believes that it is about to rain.
>
> H_3. There is something that Jane's believing that she believes that it is about to rain is wholly constituted by and which Jane's believing that it is about to rain, her understanding and having an interest in the question of whether she believes that it is about to rain, and her being fully rational is partly constituted by.

[7] See Stoljar (2019) for a discussion of further ways of refining such claims. For the purposes of the argument, all that matters is that there is some non-trivial, finite, mental condition which necessitates the introspective belief.

Of these explanations, H_1 certainly looks to be the best. There is no candidate for the kind of third-factor required by H_3, and, at least initially, it is hard to see why Jane's believing that it is about to rain, her understanding and taking an interest in the question of whether she believes that it is about to rain, and her being fully rational, would be partly constituted by her believing that she believes that it is about to rain: none of these conditions seem to be independently partly constituted by her believing that she believes that it is about to rain, and it is hard to see how, jointly, they could be partly constituted by her believing that she believes that it is about to rain. So we may tentatively conclude that H_1 provides the best explanation of the necessary connection.

These two steps, then, provide motivation for the two premises of the argument from necessitation for constitutivism about introspective belief. We can think of the argument as proceeding as follows:

P_1 There is a necessary connection between (i) Jane's believing that she believes that it is about to rain and (ii) her believing that it is about to rain, her understanding and having some interest in the question of whether she believes that it is about to rain, and her being fully rational.

P_2 The best explanation of this necessary connection is that Jane's believing that she believes that it is about to rain is wholly constituted by her believing that it is about to rain, her understanding and having some interest in the question of whether she believes that it is about to rain, and her being fully rational.

C_1 Jane's believing that she believes that it is about to rain is wholly constituted by her believing that it is about to rain, her understanding and having some interest in the question of whether she believes that it is about to rain, and her being fully rational.

Since the example we have focused on here was entirely arbitrary, the same reasoning can be followed in arguing for constitutivism about introspective belief as a general thesis. At least initially, then, the argument from necessitation provides a good case for constitutivism about introspective belief.

2.3 *Evaluating the Arguments*

We can now turn to the evaluation of the arguments from necessitation. Whether the arguments are successful turns crucially on how the notion of rationality is understood. So far I have presented the arguments without comment on how rationality is to be understood. I will now argue that there are two understandings of "being fully rational" which are relevant to the arguments from necessitation: an evaluative sense and a dispositional sense. When the arguments are understood in terms of the former, their first premises are true, but their second premises are false. When the arguments are understood in terms of the latter, their first premises are false. The arguments get whatever force they have from equivocating on these two understandings of "being fully rational." This objection to the arguments from necessitation is due, in its essentials, to Amy Kind, who makes the objection in connection with the argument from necessitation for constitutivism about introspective belief (2003).[8] While the objection has been generally appreciated in the philosophical literature on introspection (Gertler 2010), it has not been generally appreciated in the philosophical literature on instrumental desires.[9]

It is natural to think about the arguments above on an evaluative understanding of "being fully rational." On this understanding, someone is fully rational if and only if they are not in violation of the principles of rationality, that is, if and only if they fully conform to the principles of rationality. On this understanding, someone is less than fully rational if they do not fully conform to the principles of rationality. It is very plausible that there is a principle of rationality which requires you to desire the means if you desire some end and believe that the means are a means to that end. Similarly, it is very plausible that there is a principle of rationality which requires you to believe that you have some attitude if you have that attitude and you understand and have some interest in the question of whether you have it. This can be brought out by reflecting on the necessitation claims with this understanding of "being fully rational" made fully explicit:

> N_1'. Necessarily, if Jane desires to keep her health, believes that exercising is a means to keeping her health, and fully conforms to the principles of rationality, then she desires to exercise.
>
> N_2'. Necessarily, if Jane believes that it is about to rain, understands and has some interest in the question of whether she believes that

[8] While the objection, in its essentials, is due to Kind, the specific development of the objection made here is original to this essay.

[9] In both Kind's (2003) and Gertler's (2010) discussions, there is an appeal to causation, and a contrast between causation and constitution, which is not made in the presentation of the objection here.

doi: 10.48106/dial.v74.i4.02

> it is about to rain, and fully conforms to the principles of rationality, then she believes that she believes that it is about to rain.

Since it is arguably this understanding of "being fully rational" which we evaluated the original necessitation claims with, it is not surprising that they both appear to be plausible when this understanding is made fully explicit.

The problem arises for the arguments from necessitation when we turn to the evaluation of their second premises on this understanding. To see the problem in the case of instrumental desire, notice that if Jane desires to keep her health and believes that exercising is a means to keeping her health, then, in order to fully conform to the principles of rationality, she must desire to exercise. But then, given that she desires to keep her health and believes that exercising is a means to keeping her health, *if* she fully conforms to the principles of rationality, this must be at least partly because she desires to exercise. So, her fully conforming to the principles of rationality is partly constituted by her desiring to exercise. And if her fully conforming to the principles of rationality is partly constituted by her desiring to exercise, her desiring to exercise cannot be even partly constituted by her conforming to the principles of rationality. So, on this understanding of "being fully rational," H_1 is not the best explanation of the necessary connection and the argument from necessitation fails.

To see the problem in the case of introspective belief, notice that if Jane believes that it is about to rain, and understands and has some interest in the question of whether she believes that it is about to rain, then, in order to fully conform to the principles of rationality, she must believe that she believes that it is about to rain. But then, given that she believes that it is about to rain, and understands and has some interest in the question of whether she believes that it is about to rain, *if* she fully conforms to the principles of rationality, this must be at least partly because she believes that she believes that it is about to rain. So, her fully conforming to the principles of rationality is partly constituted by her believing that she believes that it is about to rain. And if her fully conforming to the principles of rationality is partly constituted by her believing that she believes that it is about to rain, her believing that she believes that it is about to rain cannot be even partly constituted by her fully conforming to the principles of rationality. So on this understanding H_1 is not the best explanation of the necessary connection and the argument from necessitation fails.

An analogy might help to drive home the crucial point here. Suppose that you are legally required to pay your taxes before the end of the financial year every financial year. Then, you fully conform to the law only if you pay your taxes before the end of the financial year every year. Now, if it is the end of the financial year, then *if* you fully conform to the law, this must be at least partly because you have paid your taxes. So, your fully conforming to the law is partly constituted by your paying your taxes. And if your fully conforming to the law is partly constituted by your paying your taxes, then your paying your taxes cannot be, even partly, constituted by your fully conforming to the law. As this analogy demonstrates, while conforming to certain norms may, along with other conditions, necessitate certain further conditions, it is your conforming to the norms which is partly constituted by those further conditions, and not vice versa. It is not surprising then that on the evaluative understanding of "being fully rational" the second premises of the arguments from necessitation are false.

At this stage a proponent of the arguments from necessitation may argue that this is not the understanding of "being fully rational" they had in mind. They may instead appeal to a dispositional understanding of "being fully rational." On this understanding, being fully rational is being disposed to conform to the principles of rationality (perhaps along with there being no barrier to one's manifesting this disposition). Someone who was fully disposed to conform to the principles of rationality, where there is no barrier to their manifesting this disposition, would very plausibly *come to desire* to exercise if they desired to keep their health and believed that exercising was a means to keeping their health. Someone who was fully disposed to conform to the principles of rationality, where there was no barrier to their manifesting this disposition, would very plausibly *come to believe* that they believed that it was about to rain if they believed that it was about to rain and understood and had some interest in the question of whether they believed that it was about to rain. The trouble with this understanding is that the first premises of the arguments from necessitation seem to be false if they are understood in terms of it. To see the trouble, we can reflect on the necessitation claims with this understanding of "being fully rational" made fully explicit:

> N_1''. Necessarily, if Jane desires to keep her health, believes that exercising is a means to keeping her health, is disposed to conform to the principles of rationality, and there is no barrier to the manifestation of her disposition, then she desires to exercise.

doi: 10.48106/dial.v74.i4.02

> N_2''. Necessarily, if Jane believes that it is about to rain, understands and has some interest in the question of whether she believes that it is about to rain, is disposed to conform to the principles of rationality, and there is no barrier to the manifestation of her disposition, then she believes that she believes that it is about to rain.

The problem here is that even when all barriers to the manifestation of a disposition are removed, there is no metaphysically necessary connection between having the disposition, the triggering conditions for the disposition obtaining, and the disposition manifesting. At best, the relation between having a disposition, the triggering conditions for that disposition obtaining, and the manifestation of the disposition, is one of *nomological necessity*. On this understanding of "being fully rational" the first premises of the arguments from necessitation are false, and there is no necessary connection between being in certain psychological conditions, being fully rational, and being in some further psychological condition.

It might be thought at this point that we might add some further condition to the antecedents of the conditionals above to avoid this problem. My response to this suggestion, however, is that there is nothing that can be added that will not either (i) fall prey to considerations like those just given against the current proposal or (ii) fall prey to considerations like those given against the evaluative understanding of "being fully rational." Suppose someone suggests that we need only add to the antecedents of the conditionals that the dispositions be manifested. Then we would have:

> N_1'''. Necessarily, if Jane desires to keep her health, believes that exercising is a means to keeping her health, is disposed to conform to the principles of rationality, and manifests this disposition, then she desires to exercise.

> N_2'''. Necessarily, if Jane believes that it is about to rain, and has some interest in the question of whether she believes that it is about to rain, is disposed to conform to the principles of rationality, and manifests this disposition, then she believes that she believes that it is about to rain.

But this suggestion faces a problem analogous to that faced by the view which understands "being fully rational" in the evaluative sense. Given that Jane desires to keep her health, believes that exercising is a means to keeping her health, and is disposed to conform to the principles of rationality, *if* she manifests her disposition, this must be in part because she desires to exercise. To manifest this disposition is, in part, to come to desire to exercise. So, her manifesting this disposition is partly constituted by her desiring to exercise. And if her manifesting this disposition is partly constituted by her desiring to exercise, her desiring to exercise cannot be even partly constituted by her manifesting this disposition (and so cannot be wholly constituted by her desiring to keep her health, believing that exercising is a means to keeping her health, being disposed to conform to the principles of rationality, and manifesting this disposition). I suspect that Smith's appeal to the idea that a means-end belief and a non-instrumental desire must stand "in a suitable relation" in his statement of constitutivism about instrumental desire is an attempt to straddle the gap here between the triggering conditions of a disposition obtaining and the disposition manifesting. As we have just seen, however, if we understand the claim in terms of the triggering conditions being met—the means-end belief and the non-instrumental desire are appropriately related just when they trigger the relevant disposition—then Smith's position falls prey to considerations like those given against the dispositional understanding, and if we understand it in terms of the disposition manifesting—the means-end belief and the non-instrumental desire are appropriately related just when the relevant disposition manifests—then it falls prey to the considerations just given.

As initially compelling as the arguments from necessitation may seem, they are ultimately unsuccessful. As I said earlier, this is old news in the literature on introspective belief. There, it is widely conceded that Sydney Shoemaker may be right about the necessitation claim, at least if it is understood on the evaluative sense of "rational," but it is held that nothing follows from this vis-à-vis the constitutive theory of introspective belief. The problem is that Shoemaker moves too quickly from the necessitation claim to the constitution claim. To see this, consider the passage quoted earlier from Shoemaker. There Shoemaker moves from a supervenience claim in the first sentence to the constitution claim in the second. Indeed, he says that he means the same thing by both claims. But, at best, he has only argued for the metaphysical

doi: 10.48106/dial.v74.i4.02

necessitation or supervenience claim, and the constitution claim does not immediately follow, as we have seen.[10]

It is doubtful that Shoemaker had the evaluative sense of "rational" in mind, however. In one of the only places where Shoemaker gives us any clues about the sense of rationality he has in mind, he writes: "The fact that the person is rational might be compared to the fact that the powder in the bomb was dry" (1996, 32). This certainly suggests that Shoemaker had the dispositional sense in mind. And, as we have seen, it is plausible that *if* the necessitation claim were true on the dispositional sense of "rational," then the constitution claim would follow. It is widely agreed, however, that Shoemaker's arguments for the necessitation claim do not establish the necessitation claim. Those who think that the necessitation claim is true, think that it is true on the evaluative sense of "rational," but for independent reasons.[11] I suspect that no argument could establish the necessitation claim on the dispositional sense of "rational," so I am sceptical about the prospects of the argument from necessitation. Nonetheless, as I will now argue, there is a better argument for constitutivism about introspective belief available, one which is implicit in Shoemaker's work but which gets overshadowed by the argument from necessitation.

3 The Arguments from Cognitive Dynamics

In this section, I will formulate and evaluate the argument from cognitive dynamics for constitutivism about instrumental desire and the argument from cognitive dynamics for constitutivism about introspective belief. I will argue that the former is a good argument for constitutivism about instrumental desire, and, in light of the conclusions of the previous section, the *best* argument for this thesis. Then I will argue that the latter is a good argument, if not an equally good argument, for constitutivism about introspective belief.

3.1 *For Constitutivism About Instrumental Desire*

The argument from cognitive dynamics for constitutivism about instrumental desire begins from an observation about the cognitive dynamics of instrumen-

[10] The supervenience claim is the conclusion of Shoemaker's famous argument from self-blindness. See Shoemaker (1996, 47–48).

[11] Kieran Setiya argues that a nearby necessitation claim is true, and that this is what Shoemaker has correctly drawn attention to (Setiya 2011). But what gets necessitated, according to Setiya, is a capacity for introspective belief, not introspective belief itself.

tal desires. The observation is that they systematically come and go with the desires and means-end beliefs which explain them.

> D_1. (i) If someone comes to desire to Ψ and to believe that Φ-ing is a means to Ψ-ing, they will come to desire to Φ if they are fully rational and (ii) if someone merely instrumentally desires to Φ and they cease desiring to Ψ or cease believing that Φ-ing is a means to Ψ-ing, then they will cease desiring to Φ if they are fully rational.

These are claims about the cognitive dynamics of particular desires. They are the cognitive dynamics of someone who is rational, that is, someone who is disposed to conform to the principles of rationality. These cognitive dynamics are partly constitutive of what it is to be disposed to conform to the principles of rationality. It is because we observe these cognitive dynamics that we believe that we are rational in this sense.

How could a cognitive system exhibit these dynamics? To see how, let's assume a broadly functionalist picture of beliefs and desires. On this picture, for X to believe that P is for X to be in some state or other which plays the believing-that-P-role, and for X to desire to Φ is for X to be in some state or other which plays the desiring-to-Φ-role. We could then re-describe the cognitive dynamics above in terms of ceasing to be in a state which plays the desiring-to-Φ-role under certain conditions, and coming to be in a state which plays the desiring-to-Φ-role under certain conditions. But now two importantly different hypotheses arise concerning the relations between these states.

According to one hypothesis, call it the *causal hypothesis*, the dynamics are explained by the fact that when someone comes to be in a state which plays the desiring-to-Φ-role, and someone comes to be in a state which plays the believing-that-Ψ-ing-is-a-means-to-Φ-ing-role, their coming to be in these states jointly *causes* them—by means of their manifesting a rational disposition—to come to be in a state which plays the desiring-to-Ψ-role. Similarly, someone's ceasing to be in a state which plays the believing-that-Ψ-ing-is-a-means-to-Φ-ing-role or ceasing to be in a state which plays the desiring-to-Φ-role will cause them to cease being in a state which plays the desiring-to-Ψ-role. Rational dispositions can then be thought of as ordinary causal dispositions, where the triggering conditions are thought of as causes of the manifestations. On this hypothesis the cognitive dynamics are explained by various causal transactions between constitutively distinct states or events involving consti-

doi: 10.48106/dial.v74.i4.02

tutively distinct states (they must be constitutively distinct in order to stand in causal relations).

According to another hypothesis, call it the *constitutive hypothesis*, the dynamics are explained by the fact that when someone comes to be in a state which plays the desiring-to-Φ-role and someone comes to be in a state which plays the believing-that-Ψ-ing-is-a-means-to-Φ-ing-role, their coming to be in these states constitutes—by means of their manifesting a rational disposition—their coming to be in a state which plays the desiring-to-Ψ-role. It does so because the former states together *constitute* a state which plays the desiring-to-Φ-role. Similarly, someone's ceasing to be in a state which plays the believing-that-Ψ-ing-is-a-means-to-Φ-ing-role or ceasing to be in a state which plays the desiring-to-Φ-role just is their ceasing to be in a state which plays the desiring-to-Ψ-role. Rational dispositions, on this hypothesis, are not ordinary causal dispositions. They are what we might call constitutive dispositions, since the triggering conditions bear a constitutive relation to the manifestations. On this hypothesis the cognitive dynamics are explained in terms of states which play particular roles jointly constituting states which play other roles. The relevant thing about the constitutive hypothesis is that, if it is true, then instrumental desires are not distinct from the desires and means-end beliefs which explain them. This is because instrumental desires are partly constituted by states which partly constitute the corresponding desires and means-end beliefs which explain them. So if there is an argument for the constitutive hypothesis, there is an argument for constitutivism about instrumental desire.

Each of these hypotheses is clearly an empirical hypothesis. If cognitive science were so advanced that we could determine which states play which roles, then, in principle we could settle the question of whether the states are constitutively distinct and causally related or constitutively non-distinct and constitutively rather than causally related. But we are far from being able to answer the question this way. The best we have, and the best we may ever have, is indirect evidence for one hypothesis over the other based on arguments to the best explanation of the observed cognitive dynamics. Let's consider the evidence for and against, then.

Perhaps the weakest consideration in favour of the constitutive hypothesis, but a consideration nonetheless, comes from the relative cognitive efficiency of having the states which play the role of certain desires being constituted by the states which play the roles of other desires and means-end beliefs, rather than having the former be distinct from and caused by the latter. A cognitive system requires far fewer distinct states and fewer dependencies between them in order to have a wide range of instrumental desires on the constitutive hypothesis. To put the point in slogan form: *the constitutive hypothesis is cognitively more efficient than the causal hypothesis.*

A stronger consideration in favour of the constitutive hypothesis comes from the observation that the causal roles which are definitive of desires and means-end beliefs, along with the principles of rationality, predict that by merely having those desires and means-end beliefs, and being rational, the agent will be disposed to act *as if* she desired the means. If desiring the means were a matter of coming to be in a distinct state which plays the role of desiring the means, as the causal hypothesis holds, then the disposition to act as if one desired the means would be over-determined. There's nothing by way of the agent's dispositions to act that being in this state would contribute which is not already contributed by their being in these other states and their being rational. The state is motivationally redundant.[12] So, to put the point in slogan form: *the constitutive hypothesis avoids the prediction that instrumental desires are motivationally redundant.*

Perhaps the strongest consideration in favour of the constitutive hypothesis, the one that I am willing to put the most weight on, begins with an observation about the strength of the dependence of instrumental desires on means-end beliefs and other desires. As we saw earlier, one claim about the cognitive dynamics of instrumental desires is that if someone merely instrumentally desires to Φ and they cease desiring to Ψ or cease believing that Φ-ing is a means to Ψ-ing, then they will cease desiring to Φ. While this claim is compatible with both the constitutive and causal hypotheses, the constitutive hypothesis has a far better explanation of it. Indeed, the explanation comes for free on the constitutive hypothesis, since it is no surprise that when one ceases to be in either of the states which jointly constitutes the state which plays the desiring-to-Φ-role that one will cease desiring to Φ. The causal hypothesis requires the auxiliary hypothesis here that when the state which plays the desiring-to-Φ-role is caused and explained by the states which play the other roles, it will remain causally dependent on those states. Of course, it is possible that, simply as a matter of fact, such a state will remain causally dependent on these other states. But this claim has to be added as an auxiliary hypothesis to the causal hypothesis, thus making the hypothesis more complicated than the constitutive hypothesis. And, moreover, there is every reason to think

[12] See Arpaly and Schroeder (2014, 9) for a similar observation.

that this causal dependence would sometimes break down, giving rise to stray instrumental desires, desires which are no longer dependent on the means-end beliefs and desires which caused them in the first place. But if this phenomenon exists in our psychology, it remains unobserved. The constitutive hypothesis correctly predicts that there will be no stray instrumental desires. To sum up the point of this paragraph in slogan form: *instrumental desires are deeply dependent on other mental states for their existence and the constitutive hypothesis best explains this*.

It is basically this consideration which motivates the argument in the following passage from Smith:

> It is a striking fact that instrumental desires disappear immediately an agent loses either the relevant non-instrumental desire or means-end belief [...]. Yet there is no reason why this should be so if an instrumental desire were merely a desire that has a non-instrumental desire and a means-end belief somewhere in its causal history. Why should a desire disappear when (say) the desire that caused it, way back when, disappears? Instrumental desires are thus better thought of as being nothing over and above the non-instrumental desires and means-end beliefs that explain them. (2004, 96)

Smith begins here with the observation that instrumental desires are deeply dependent on the desires and means-end beliefs which explain them. He then argues against the causal hypothesis and for the constitutive hypothesis on the basis of the fact that the latter provides a better explanation of the observation than the former. Of course, Smith's argument against the causal hypothesis is too fast. An instrumental desire could remain causally dependent on another desire and a means-end belief in the way that a light's being on remains causally dependent on the light switch's being turned on. It could be that the relation between an instrumental desire and the desire and means-end belief which explains it is like this. But the constitutive hypothesis nonetheless provides a better explanation of the deep dependence between an instrumental desire and the desire and means-end belief which explains it.

The best argument, then, for constitutivism about instrumental desire is the argument from cognitive dynamics. Unlike the argument from necessitation for constitutivism about instrumental desire, which would decisively establish constitutivism about instrumental desire if it were cogent, the argument from cognitive dynamics makes constitutivism about instrumental desire the conclusion of an ordinary argument to the best explanation. But it isn't really surprising that this should be so, since constitutivism about instrumental desire is most plausibly thought of as a contingent hypothesis about instrumental desires.

3.2 *For Constitutivism About Introspective Belief*

If the argument from cognitive dynamics for constitutivism about instrumental desire provides good support for constitutivism about instrumental desire, then perhaps an analogous argument from cognitive dynamics could provide good support for constitutivism about introspective belief. In this section I argue that it does.

The argument from cognitive dynamics for constitutivism about introspective belief begins from an observation about the cognitive dynamics of introspective beliefs. The observation is that introspective beliefs systematically come and go with the mental states—or in the case under consideration, the attitudes—that they are about.

> D_2. (i) If someone has some interest in the question of whether they Ψ that P, and they come to Ψ that P[13], then they will come to believe that they Ψ that P if they are fully rational, and (ii) if someone introspectively believes that they Ψ that P, and they cease Ψ-ing that P, then they will cease believing that they Ψ that P if they are fully rational.

These are claims about the cognitive dynamics of particular beliefs. They are the cognitive dynamics of someone who is rational, that is, someone who is disposed to conform to the principles of rationality. These cognitive dynamics are partly constitutive of what it is to be disposed to conform to the principles of rationality. It is because we observe these cognitive dynamics that we believe that we are rational in this sense.

Now, let me be upfront here about an important disanalogy with the observation about the cognitive dynamics of instrumental desires. The observation just given is likely to strike many as highly controversial. Failures of introspective belief are the norm. Failures of instrumental desire are the exception. I have two responses to this kind of pessimism about introspective belief. The

[13] The antecedent of this conditional may need strengthening. Perhaps one needs to consider the question of whether one Ψ-s that P also.

doi: 10.48106/dial.v74.i4.02

first is that it vastly overstates the case. To say that failures of introspective belief are the norm is to overlook the wide range of cases where introspective belief is utterly unproblematic. Over a vast range of mundane beliefs, desires, and other mental states, I have utterly unproblematic introspective access. If failures of introspective belief were the norm here, our mental life would be in serious trouble. The second response is that nothing in the argument from cognitive dynamics depends on an overly optimistic view of our capacity for introspective belief. It may well be that failures of local rationality are far more common in the case of introspective beliefs, but as long as introspective beliefs have some of the features I draw attention to below, the argument from cognitive dynamics will go through. This is a point I will return to after presenting the rest of the argument.

How could a cognitive system exhibit the dynamics above? Again, assuming a broadly functionalist picture of beliefs and desires, we can recast the observation in the following terms. When someone comes to be in a state which plays the Ψ-ing-that-P-role, they come to be in a state which plays the believing-that-one-Ψ-s-that-P-/-to-Φ-role, insofar as they are rational and have some interest in the question of whether they Ψ that P. When someone ceases to be in a state which plays the Ψ-ing-that-P-role, they cease to be in a state which plays the believing-that-one-Ψ-s-that-P-role. Again, two importantly different hypotheses arise concerning the identity of these states.

According to one hypothesis, call it the *causal hypothesis*, the dynamics are explained by the fact that when someone comes to be in a state which plays the Ψ-ing-that-P-role, they are caused to come to be in a state which plays the believing-that-one-Ψ-s-that-P-role, insofar as they are rational and have some interest in the question of whether they Ψ-that-P. And someone ceases to be in a state which plays the Ψ-ing-that-P-role, they are caused to cease to be in a state which plays the believing-that-one-Ψ-s-that-P-role. On this hypothesis the cognitive dynamics are explained by various causal transactions between constitutively distinct states or events involving these states.

According to another hypothesis, call it the *constitutive hypothesis*, the dynamics are explained by the fact that when someone comes to be in a state which plays the Ψ-ing-that-P-role, they thereby come to be in a state which plays the believing-that-one-Ψ-s-that-P-role, since the former state plays the latter role. And when someone ceases to be in a state which plays the Ψ-ing-that-P-role, they thereby cease being in a state which plays the believing-that-one-Ψ-s-that-P-role, since it was the former state which played the latter role. On this hypothesis the cognitive dynamics are explained by the states which play the first-order roles playing the second-order roles.[14]

Each of these hypotheses is clearly an empirical hypothesis. What arguments can be given for and against? Not surprisingly, the considerations are perfectly analogous to those given in the argument from cognitive dynamics for constitutivism about instrumental desire.

Perhaps the weakest consideration in favour of the constitutive hypothesis, but a consideration nonetheless, comes from the relative cognitive efficiency of having the states which play the roles of particular mental states also play the role of beliefs about those mental states, rather than having the former be distinct from and caused by the latter. A cognitive system requires far fewer distinct states and fewer dependencies between them in order to have a wide range of introspective beliefs on the constitutive hypothesis. This consideration has considerable bite in contemporary contexts where doubt has arisen, both on the basis of philosophical and empirical enquiry, concerning the claim that we have a distinct perception-like capacity for inner-sense, one which causally detects our mental states and outputs introspective beliefs. This view goes hand in hand with the causal hypothesis. The constitutive hypothesis cuts out the middle-man, and requires no distinct perception-like capacity for inner-sense.[15] To sum up the points of this paragraph in slogan form: *the constitutive hypothesis is cognitively more efficient than the causal hypothesis*.

A stronger consideration in favour of the constitutive hypothesis comes from the observation that the causal roles which are definitive of many mental states, along with the principles of rationality, predict that by merely being in those mental states, and being rational, an agent will be disposed to act *as if* she believes that she is in those mental states.[16] In particular, it has been observed that if you are in pain, say, and you are rational, you will be disposed, partly in virtue of the fact that you are in pain, to say "I am in pain."

[14] See Shoemaker (1996, 33–34, 242–244).

[15] This might be a little bit unfair to the causal hypothesis. Since there is a version of the causal hypothesis which cuts out the middle-man too, and requires no distinct perception-like capacity for inner-sense. My point here is that once we are on the lookout for cognitively efficient hypotheses about introspective beliefs, the constitutive hypothesis wins hands down.

[16] This point is well made by Shoemaker (1996). While Shoemaker makes the point in the context of arguing for the necessitation or supervenience claim, I am here making it in the context of the cognitive dynamics argument. Shoemaker sometimes says things which suggest that he might have something like the cognitive dynamics argument in mind. This is a point I will come to in the text.

doi: 10.48106/dial.v74.i4.02

This point has been made over and over by expressivists in the philosophy of introspection, who take it to show that statements like “I am in pain” do not report mental states but merely express them. But we needn’t understand the claim in this manner, since it is possible that such statements both report and express mental states, because the mental states they express double, according to the constitutive hypothesis, as beliefs about those mental states. If introspectively believing that you are in a particular mental state were a matter of coming to be in a distinct state which plays the role of a belief that you are in some mental state, as the causal hypothesis holds, then the disposition to act as if you believed that you were in that mental state would be over-determined. There’s nothing by way of the agent’s dispositions to act that being in this distinct state would contribute which is not already contributed by their being in the first-order state, their having an interest in the question of whether they are in that state, and their being fully rational. The state is motivationally redundant.[17] So, to put the point in slogan form: *the constitutive hypothesis avoids the prediction that introspective beliefs are motivationally redundant.*

Perhaps the strongest consideration in favour of the constitutive hypothesis, the one that I am willing to put the most weight on, begins with an observation about the strength of the dependence of introspective beliefs on the mental states they are about. As we saw earlier, one claim about the cognitive dynamics of introspective belief is that if someone introspectively believes that they Ψ that P, and they cease Ψ-ing that P, then they will cease believing that they Ψ that P.[18] While this claim is compatible with both the constitutive and causal hypotheses, the constitutive hypothesis has a far better explanation of it. Indeed, the explanation comes for free on the constitutive hypothesis, since it is no surprise that when one ceases to be in the state which plays the believing-that-one-Ψ-s-that-P-role, one will cease believing that one Ψs that P. The causal hypothesis requires the auxiliary hypothesis here that when the state which plays the believing-that-one-Ψ-s-that-P-role is caused and explained by the state which plays the Ψ-that-P-role, it will remain causally

[17] To be clear: I am not denying that introspective beliefs themselves are motivationally redundant, that they make no difference to the cognitive functioning of the mental states they are about; I am only claiming that a distinct state which had these consequences would be redundant insofar as being in the mental states in question, having an interest in the question of whether you are in those mental states, and being fully rational would already have the consequences for one’s cognitive life we take introspective beliefs to have.

[18] This was pointed out to me by Daniel Nolan.

dependent on this state. Of course, it is possible that, simply as a matter of fact, such a state will remain causally dependent on this other state. But this claim has to be added as an auxiliary hypothesis to the causal hypothesis, thus making the hypothesis more complicated than the constitutive hypothesis. And, moreover, there is every reason to think that this causal dependence would sometimes break down, giving rise to stray introspective beliefs, *introspective* beliefs which are no longer dependent on the mental states which caused them in the first place. But if this phenomenon exists in our psychology, it remains unobserved. The constitutive hypothesis correctly predicts that there will be no stray introspective beliefs. And this is a significant point in its favour. So, to sum up the points of this paragraph in slogan form: *introspective beliefs are deeply dependent on other mental states for their existence and the constitutive hypothesis best explains this*.

While Shoemaker is more closely associated with the argument from necessitation, it is clear that he also has something like the argument from cognitive dynamics in mind. Indeed, I think that, to the extent that he does have the latter in mind, this is the best argument he has for constitutivism about introspective belief. Consider the following passage from a recent paper of Shoemaker's defending constitutivism about introspective belief:

> One might, indeed, wonder whether there is any need to postulate standing second-order beliefs that self-ascribe available first-order beliefs. It goes with having the available first-order belief that *p* that if the question whether one believes that *p* arises, one will judge that one does—one will assent to the proposition that one believes that *p*. But this seems to be the result of one's having the belief that *p*, not the result of one's having a second-order belief whose cognitive dynamics is independent of that of the belief that *p*, in the way that the cognitive dynamics of one's belief about another person's belief is independent of that of the other person's belief. It would seem inefficient for our psychology to involve the storage of standing second-order beliefs ascribing available first-order beliefs, if there is nothing for these second-order beliefs to do that is not done by the first-order beliefs themselves. (2012, 247)

These remarks combine elements of all three of the considerations I have given above. Shoemaker makes a claim about efficiency, there is also a claim about redundancy, and he speaks of the predicted cognitive independence of introspective beliefs from the mental states they are about on the causal hypothesis. I have teased out these considerations and argued that together they add up to a reasonable case for the constitutive hypothesis.

The best argument, then, for constitutivism about introspective belief is the argument from cognitive dynamics. Unlike the argument from necessitation for constitutivism about introspective belief, which would decisively establish constitutivism about introspective belief if it were cogent, the argument from cognitive dynamics makes constitutivism about introspective belief the conclusion of an ordinary argument to the best explanation. But it isn't really surprising that this should be so, since constitutivism about introspective belief is most plausibly thought of as a contingent hypothesis about introspective beliefs.

To end, let me return to the obvious line of criticism which may be raised against the argument from cognitive dynamics for constitutivism about introspective belief. The criticism is basically that it depends on far too rosy a picture of introspective belief. But we are now in a position to see that it does not. The consideration about efficiency requires only that we have a significant number of introspective beliefs, so that considerations of efficiency come into play. It doesn't require that we approximate omniscience and infallibility. The considerations about redundancy, likewise, only require that we have a significant number of introspective beliefs, so that considerations about redundancy come into play. And, finally, considerations about deep dependence do not require that we are rarely in error about our own mental states. This is an important point. We may have many false beliefs about our own mental states. We may often be in error about our own mental states. But as long as those beliefs about our own mental states arrived at by introspection remain deeply dependent on the mental states they are about—that is, as long as there are no stray *introspective* beliefs—the point about dependence holds.[*]

Ryan Cox
0000-0002-1381-448X
The University of Sydney

[*] Thanks to Michael Smith for discussion on a very early version of this paper. Thanks also to members of the ANU Philosophy of Mind Work in Progress Group for their helpful feedback on an earlier version of this paper. Finally, thanks to several reviewers for *Dialectica* for their helpful comments.

ryan.cox@sydney.edu.au

References

ARMSTRONG, David M. 1968. *A Materialist Theory of the Mind*. London: Routledge & Kegan Paul.

ARPALY, Nomy and SCHROEDER, Timothy. 2014. *In Praise of Desire*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199348169.001.0001.

BOYLE, Matthew. 2009. "Two Kinds of Self-Knowledge." *Philosophy and Phenomenological Research* 78(1): 133–164, doi:10.1111/j.1933-1592.2008.00235.x.

BYRNE, Alex. 2011. "Transparency, Belief, Intention." *Proceedings of the Aristotelian Society, Supplementary Volume* 85: 201–221, doi:10.1111/j.1467-8349.2011.00203.x.

CASSAM, Quassim. 2014. *Self-Knowledge for Humans*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199657575.001.0001.

CHALMERS, David J. 2010. *The Character of Consciousness*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780195311105.001.0001.

DAVIS, Wayne A. 1986. "The Two Senses of Desire." in *The Ways of Desire: New Essays in Philosophical Psychology on the Concept of Wanting*, edited by Joel MARKS, pp. 63–82. Chicago, Illinois: Precedent Publishing, Inc.

GERTLER, Brie. 2010. *Self-Knowledge*. London: Routledge.

GORDON, Robert M. 1986. "Folk Psychology as Simulation." *Mind and Language* 1(2): 158–171, doi:10.1111/j.1468-0017.1986.tb00324.x.

HUME, David. 1975. *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*. 3rd ed. Oxford: Oxford University Press. Edited by Lewis Amherst Selby-Bigge; revised and with notes by Peter Harold Nidditch.

KIND, Amy. 2003. "Shoemaker, Self-Blindness and Moore's Paradox." *The Philosophical Quarterly* 53(210): 39–48, doi:10.1111/1467-9213.00294.

MARKS, Joel. 1986. "Introduction: On the Need for a Theory of Desire." in *The Ways of Desire: New Essays in Philosophical Psychology on the Concept of Wanting*, edited by Joel MARKS, pp. 1–15. Chicago, Illinois: Precedent Publishing, Inc.

MCDANIEL, Kris and BRADLEY, Ben. 2008. "Desires." *Mind* 117(466): 267–302, doi:10.1093/mind/fzn044.

MORAN, Richard. 2001. *Authority and Estrangement. An Essay on Self-Knowledge*. Princeton, New Jersey: Princeton University Press.

SCHROEDER, Timothy. 2004. "Functions from Regulation." *The Monist* 87(1): 115–135, doi:10.5840/monist20048717.

SETIYA, Kieran. 2011. "Knowledge of Intention." in *Essays on Anscombe's* Intention, edited by Anton FORD, Jennifer HORNSBY, and Frederick STOUTLAND, pp. 170–197. Cambridge, Massachusetts: Harvard University Press.

SHOEMAKER, Sydney S. 1996. *The First-Person Perspective and Other Essays*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511624674.

—. 2012. "Self-Intimation and Second-Order Belief." in *Introspection and Consciousness*, edited by Declan Smithies and Daniel Stoljar, pp. 239–258. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199744794.001.0001.

Smith, Michael A. 2004. "Instrumental Desires, Instrumental Rationality." *Proceedings of the Aristotelian Society, Supplementary Volume* 78: 93–109, doi:10.1111/j.0309-7013.2004.00117.x.

Stoljar, Daniel. 2019. "Evans on Transparency: A Rationalist Account." *Philosophical Studies* 176(8): 2067–2085, doi:10.1007/s11098-018-1111-x.

The Mental States First Theory of Promising

Alida Liberman

Most theories of promising are insufficiently broad, for they ground promissory obligation in some external or contingent feature of the promise. In this paper, I introduce a new kind of theory. The Mental States First (MSF) theory grounds promissory obligation in something internal and essential: the mental state expressed by promising, or the state that promisors purport to be in. My defense of MSF relies on three claims. First, promising to Φ expresses that you have resolved to Φ. Second, resolving to Φ commits you to Φing, all else being equal. Third, the norms on speech acts are determined by the norms on the mental states they express, such that publicly expressing that you are in a state subjects you to whatever commitments are normally incurred by being in that state, regardless of whether you really are in it. I suggest that this general approach might also explain how the norms on other sorts of speech acts work.

Philosophers have offered a variety of theories of promissory obligation, most of which ground promissory obligation in some external or contingent feature of the promise. In this paper, I sketch a new kind of theory, which instead grounds promissory obligation in something internal and essential to promise-making. This *Mental States First* theory of promising (or MSF) posits that the norms on promises are fixed or determined by the norms on the promises' underlying mental states; the mental states are "first" in the sense that they are explanatorily prior. My aim is to show that MSF should be taken seriously, because it is grounded in plausible assumptions, can accommodate a wider range of cases than can most theories, and fruitfully situates promising as part of a broader pattern of speech acts that behave in similar ways.

I motivate the search for a new theory of promissory obligation in section 1 by pointing out that most theories face cases of apparently genuine promissory obligation that they cannot accommodate. In section 2, I lay out how MSF works. In sections 3, 4, 5, I sketch arguments for several claims on which the success of MSF depends; if these claims hold, then MSF is a viable theory of promissory obligation. My argument in section 5 highlights how the general MSF pattern holds for other speech act/mental state pairs, which suggests that a Mental States First approach might be productive for understanding the norms on other kinds of speech acts, as well.

1 Motivating MSF with Marginal Cases

Most theories of promissory obligation have no trouble explaining why we are obligated to keep our promises in paradigmatic cases, or those in which the promise is something the promisee wants and expects will occur, the promisor intends to bind herself to act as promised, harm would occur were the promise broken, and the promise occurs within some obvious social convention of promising. However, these theories struggle with explaining why we are obligated to keep promises in non-paradigmatic cases, which I call *marginal cases.*[1] These are cases in which there are good theoretical arguments—or a strong intuitive presumption with no good arguments to the contrary—that the case generates a genuine promissory obligation, but in which it is doubtful whether the theory can explain or accommodate the case as an instance of promissory obligation. Proponents of particular theories tend to wield such cases as dialectical weapons against each other (e.g., by arguing that view A is inadequate because it cannot capture case X, or that view B should be preferred because it can).[2] This approach presumes that theories should be able to explain every plausible case of promissory obligation, which makes sense as a desideratum for a good account; more theoretical breadth and explanatory power is usually better than less. Moreover, since promises

[1] My use of "marginal" to describe these cases is not meant to imply that these cases only barely count as promises. Rather, I mean to convey that they are cases that ought to be considered genuine promises but that lack some of the core features of the most obvious and paradigmatic promises, which places them at the edges or borders of our practices of promising. Thanks to an anonymous referee for discussion of this point.

[2] For example, Scanlon criticizes conventionalist views in this way, by arguing that they cannot accommodate a proposed case of promising without a social practice. And Scanlon's critics make similar sorts of arguments about his view, claiming that it fails because it is subject to a counterexample; see Cholbi (2002) and Southwood and Friedrich (2009). Of course, philosophers who argue in such ways typically also offer theoretical objections against the view they are targeting. But they generally assume that being unable to accommodate a case is highly problematic for a view.

appear to be a unified phenomenon, we want a unified account of them if possible.

It is important to note that a promise does *not* count as marginal for a theory simply because the theory implies that there are circumstances under which breaking the promise is morally permissible, all-things-considered. For it is widely accepted that promises yield pro tanto moral obligations, which can be overridden if excusing conditions arise (such as needing to break the promise in order to satisfy a more important conflicting obligation). When I claim that a case is marginal for a theory, I am claiming that the theory struggles to explain why the promise in the case has *any* moral force, or why there is even a pro tanto obligation to keep it.

In the rest of this section, I give a brief overview of how the most popular theories of promissory obligation are subject to marginal cases.[3] We can sort views with contemporary traction into three broad categories.[4] First are conventionalist accounts like that of David Hume—according to which we have reason to keep promises because it is bad for us (and our reputations as trustworthy people) if we do not—or John Rawls, who argues that failing to keep promises problematically free-rides on a valuable social practice.[5] Second are expectationalist accounts like that of T.M. Scanlon, according to which promise-breaking is impermissible because it involves violating expectations that you have raised in the promisee.[6] A third class of recently popular

[3] This is not meant to be an exhaustive survey of theories of promissory obligation. I do not address views without much popular support or influence in the contemporary literature on promissory obligation, such as virtue-ethical approaches, intuitionist accounts, and Kantian deontology. Nor do I discuss act consequentialist views, according to which promises do not generate pro tanto obligations, and promise-keeping is morally required only if this leads to the best overall consequences. Of necessity, my discussions of the theories that I do address are sketchy and superficial; for more detail on how these (and other kinds of) theories of promissory obligation are subject to marginal cases, see Liberman (2015, chap. 2).

[4] This follows Heuer's (2012) helpful classification.

[5] See Hume (T 3.2.5.10) and Rawls (1971). More recent conventionalist accounts vary greatly in their details; for example, Sheinman (2008) argues that promisors can give themselves social practice-based reasons to act by communicating the intention to give themselves such a reason. Hooker (2011) offers a rule consequentialist view that can be understood as a conventionalist view, which grounds promissory obligation in the rules that would lead to the best consequences were people to internalize them (see Liberman 2020 for further discussion of this view). A number of theorists have also proposed hybrid conventionalist/expectationalist views, which ground promissory obligation in expectations that can only be generated from within an existing social practice; see Kolodny and Wallace (2003).

[6] Scanlon invokes a principle of fidelity that is justifiable on contractualist grounds; see (1990) and chapter 7 of (1998). Other views of the same sort ground promissory obligation in something

doi: 10.48106/dial.v74.i4.03

views are normative power accounts, which posit that we have the ability to change the normative situation directly and by declaration by exercising a normative power, through which we "change what someone is obliged to do by intentionally communicating the intention of hereby so doing" (Owens 2012, 4). Normative power theories claim that promissory obligation stems from an exercise of such power, and that these powers exist because they are valuable for us to have. In general, these views claim that X is an important feature of our normative lives, and that we can possess X only by having the ability to bind ourselves to each other through promising. In a transcendental step, they conclude that we therefore must possess the ability to bind ourselves to each other through promising, lest X be inaccessible.[7]

Each of these types of view—conventionalist, expectationalist, and normative power—gets something right, as each captures an important feature of paradigmatic promise-making, a feature which tells us something about why we are obligated to keep promises in many or even most cases. But each of these theories is also in some way incomplete; while they can explain why a moral obligation is present in paradigmatic cases of promising, they cannot explain the full breadth of cases in which we are obligated to keep our promises. For example, promise-breaking is often against one's self-interest because of the moral sanction it incurs, as Hume suggests. But we can easily imagine cases where this is not so; consider a traveler passing through a remote town who cons locals out of their cash, promises to pay them back, and disappears. Similarly, most societies have robust promising conventions that it would be unfair to free-ride upon, as Rawls claims. But we can imagine a successful

other than expectations; for example, Judith Jarvis Thomson (1990) argues that promise-keeping is required because the promisee is relying on the promisor, while Daniel Friedrich and Nicholas Southwood (2009) argue that promise-breaking violates the trust you have invited the promisee to place in you. Cholbi (2002) offers a contractualist view that grounds promissory obligation not in what the promisee *actually* expects, but in what she is *entitled to* expect; I argue in Liberman (2015, chap. 2) that this view cannot accommodate cases in which the promisee lacks moral standing to hold the promisor accountable.

[7] Normative power theorists cash out what the normative power to promise consists in and what valuable feature of our normative lives it supports in a variety of ways. For example, David Owens proposes a normative power grounded in what he calls our "authority interest," or the interest we have in having a certain kind of practical authority over others. By making a promise, the promisor "give[s] the promisee the right to require performance of the promisor" (2012, 144); this serves our authority interest in allowing us to determine whether another person is obligated to act as she has promised, and in allowing others to make such determinations for us. Other normative power theorists speak instead of the ability to transfer rights (Shiffrin 2008) or create exclusionary reasons (Raz 1977).

promise in a state of nature without such conventions; Scanlon offers a case of strangers from different societies on opposite sides of a river, whose hunting weapons have fallen on each other's sides and who manage to successfully promise to exchange the weapons in the absence of a shared convention of promising. Conventionalist views cannot accommodate such cases.[8]

Similarly, Scanlon is right to point out that expectations matter; we are morally obligated not to mislead people and upset their expectations, or allow them to detrimentally rely on us and fail to follow through. But sometimes the promisee does not expect the promisor to perform; maybe the promisor is notoriously bad at keeping his word. As Berislav Marušić notes, expectationalist views entail that if the promisor does "not succeed in forming expectations in the promisee [...] she will thus fail to incur promissory obligations; her promise, if it is one at all, won't be binding" (2013, 305). But making a promise that doesn't generate expectations shouldn't get the promisor off the moral hook; it would be problematic indeed if you were justified in promise-breaking because the promisee assumed you were unreliable, and accordingly did not expect you to keep your promise.

Normative power accounts face their own marginal cases. For example, many people assume that there is a distinctively promissory obligation to keep promises we make to people who die before the promise can be kept. Promises to the dead are marginal for authority transfer views like Owens's. For there must be a person to whom the authority is transferred and in whom the authority continues to reside—but once the promisee is dead, no such authority bearer exists. Normative power theorists also presume that sincerely promising to Φ requires intending to obligate yourself to Φ. But this constraint leaves out some cases, as there seem to be cases of sincere promises that do not involve forming such an intention. As Thomas Pink (2009) argues, promisees are generally concerned not with whether the promisor is morally obligated to perform the promised action, but with whether the promisor *actually will* perform the promised action. And we frequently make promises without specifically intending to obligate ourselves in daily life.[9] I take it that you

[8] More worryingly, Rawls's view fails to account for the directed nature of promissory obligation, as free-riding on a valuable social practice wrongs *all* of the participants in the practice equally, rather than wronging the promisee in particular. As Scanlon (1998) and Kolodny and Wallace (2003) both note, this is a major failure—although it is not a failure to accommodate a marginal case.

[9] This is admittedly an overly quick dismissal of a popular view. I give a more detailed argument for this claim in Liberman (2015, chap. 4). See also Liberman (2015, chap. 5) for more about the sincerity conditions on promise-making.

doi: 10.48106/dial.v74.i4.03

can promise to Φ while merely *foreseeing* that your promise will incur an obligation but without specifically *intending* to incur such an obligation. For example, I can promise to take good care of a baseball glove I borrow from my brother with the intention of putting his mind at ease about getting the glove back in good condition, while merely *recognizing* (but not intending) that this promise will obligate me. I presume that you might also sincerely promise to Φ without even foreseeing that you will Φ (e.g., if you make a conditional promise and you believe that the relevant condition is extremely unlikely to occur), or to sincerely promise to Φ while being a nihilist about whether obligations can ever be incurred. Normative power views cannot easily accommodate such cases.[10]

The three kinds of theory of promissory obligation fail to explain why we are obligated to keep our promises in all—and only—cases of seemingly genuine promissory obligation. Conventionalist and expectationalist accounts risk being both over- and under-inclusive, because the sources of obligation that they point to are not *distinctively* promissory: that is, they stem not from the content or nature of the promise itself, but from some other, contingent feature of the situation (such as the context within which the promise occurs, or the downstream effects of the promise, which may not always be as expected). While promises often rely on social conventions, they need not do so, as we saw with Scanlon's weapon-exchange example. We could also have social conventions (such as an honor code) that led people to scrupulously keep their word for fear of social repercussion or free-riding without invoking promises at all. Likewise, not all genuinely binding promises raise expectations. And we sometimes raise others' expectations without making promises to them (e.g., by putting on your coat and raising my expectation that you are about to leave the building). Normative power accounts purport to be grounded in a feature of promises as-such (i.e., the intention to obligate oneself that is expressed by a promise), but the feature they identify is not essential. Although forming

[10] Normative power theorists generally construe exercises of normative power as intentions to change the normative situation (e.g., through promising, consenting, giving, etc.). It might be possible to develop a normative power view according to which the interpersonal act of promise-making itself counts as a direct exercise of normative power, regardless of what intentions (e.g., to obligate yourself, or simply to act) accompany the act of promise-making. Such a view could get around the specific marginal case I am currently addressing. But it would require positing a brute normative power without an underlying explanation of how this normative power functions, which would not be as deeply explanatory or satisfying an account of how and why promises bind than MSF can offer. Thanks to an anonymous referee for discussion of this point.

an intention to obligate yourself and thereby transfer authority might be one way of creating a promissory obligation, it is not necessary.

This is not to say that theories of promissory obligation that are subject to marginal cases are without worth. To the contrary, maintaining valuable social practices and satisfying expectations matter morally. Considerations such as these often make it the case that a particular promise is morally obligatory to keep on independent grounds; people are often obligated to act as they have promised for reasons having to do with fairness, harm avoidance, and the like. But as already noted, such sources of obligation are neither maximally broad nor distinctively promissory. In the rest of this paper, I argue that there is another source of promissory obligation that is both broad enough to cover all cases, and distinctive of promising as such.

2 Introducing the Mental States First Theory of Promising

The existence of marginal cases motivates grounding a theory of promissory obligation not in the contingent and external circumstances of particular promises, but in something that all promises share, something that is internal and essential to the act of promising itself. One thing all promises have in common is that they involve a communicative act, usually a verbal speech act but sometimes a written or non-verbal act (for a nod or a stern look can successfully communicate a promise in the right context).[11] Whenever you perform such a communicative act, you *express* that you are in a certain mental state, by which I mean that you convey to your audience that you are in that state, and really are in it if your performance is sincere. If we can ground promissory obligation in the expression of a mental state that all promises share, we will have a maximally broad view.

This brings us to the following theory:

> Mental States First Theory of Promising (MSF). The obligation to keep promises is derived from the norms on resolutions, which are the mental state expressed by promising.

[11] Could you have promises in which no direct communication occurs? In such cases of implicit promises, there is either successful indirect communication (e.g., if Sam and Chris agree to "go steady" in a context in which this conveys a commitment to monogamy, Chris breaks a promise by dating someone else), or there is no successful communication and no promise is made (e.g., if Sam and Chris start casually dating without any mutual presumption of sexual exclusivity, Sam does not break any implicit promise by dating someone else).

doi: 10.48106/dial.v74.i4.03

MSF claims that one who promises to Φ expresses that she resolves to Φ, and *in virtue of publicly expressing that resolution* is subsequently obligated to Φ, all else being equal.[12] I am proposing that the norms on promising can be explained by appeal to the norms on the mental state expressed by promising. This is a claim about explanatory priority; the norm on the mental state explains the existence of the norm on the speech act, because it fixes or determines what that norm is.

My argument for MSF is as follows:

1. Promising to Φ expresses that you have formed a resolution to Φ, conditional on the promisee's acceptance.

[12] Downie (1985) articulates an account of promising that shares some structural similarities with MSF. Downie argues that "a promise is essentially a matter of pledging oneself" and that the promisee's "reliance and expectations are well-founded to the extent that they see that I already regard myself as obliged and they know me to be a man of my word [...]. To promise is always to state an intention in obligation-creating circumstances" (1985, 266). Downie argues that these circumstances are those in which "the intended projects have been made central and essential in one's total concerns. The self has been identified with the projects, and carrying them out has become not only a moral obligation of practical consistency but a strong moral obligation of honour or self-fidelity" (1985, 269). That is, Downie argues that there is a moral duty to maintain self-consistency about projects with which you have identified your will, and the promisor has "identified his will with the project" and "pinned his self on the future as described in his pledged actions" in a way that makes "keeping of the promise essential to the preservation of his personal integrity," such that "he will be diminished as a person if he breaks his word" (1985, 270). My view differs from Downie's in two primary ways. First, Downie's account accommodates only those promises that are deeply tied up with the promisor's sense of self, while MSF explains promissory obligation more broadly, including in cases that are not very important and that your sense of integrity is not wrapped up in. Second, Downie grounds the moral force of promissory obligation in the need to adhere to your important resolutions. While I agree that promissory resolutions generate moral obligations, I argue in Section 4.2 that this is because of the way in which they are conditional on the acceptance of the promisee, and not because there are general moral duties of self-consistency. Michael Robins (1984) offers an account that draws on action theory to ground promissory obligation in intention—specifically, in the intention that the assent of the promisee will obligate you to act as you've promised to. Robins begins with an "irreducibly normative" (1984, 12) notion of intention that binds the intending agent to act in certain ways in the future. He argues that vows are intentions about which the agent cannot change their mind, and that promises are the transferal of this "exclusionary mandate" about how one will act to the "normative control of another person" (1984, 120). Robins argues that this transfer of the exclusionary mandate to the promisee transforms the requirement to abide by one's vow into a moral obligation; various critics of Robins have argued that it is not clear what exactly this transfer consists in, how it occurs, or how it generates a moral obligation (see Cottingham 1985; Lemos 1987; Smith 1987). By contrast, MSF does not rely on any transfer (of an exclusionary mandate or any other right) to the promisee.

2. Resolving to Φ conditional on another person's acceptance rationally and morally obligates you to Φ (all else being equal).
3. In general, publicly expressing that you are in state X obligates you to act as X demands of you, regardless of whether you really are in state X.
4. Therefore, promising to Φ rationally and morally obligates you to Φ (all else being equal), regardless of whether you have really resolved to Φ.

The above argument depends on the following claims: (1) that promises express resolutions; (2) that resolutions rationally (and in some cases, morally) obligate you to act, all else being equal; and (3) the *determination claim,* according to which the norms governing the mental state expressed by a speech act (at least partially) determine what the norms on that speech act are, in both sincere and insincere cases. The latter premises are not derived from the former; these are all independent claims which together establish the conclusion. In the next three sections, I offer arguments to support each of these claims. Because I lack space to fully defend them at present, my argument is conditional: *if* these claims are true, then MSF is an appealing and viable theory of promissory obligation.

To clarify, MSF states that the mental state expressed by a promise—i.e., a resolution—is a necessary and ineliminable part of the explanation of why we are morally required to keep our promises. The mental state alone is not sufficient; an unarticulated mental state cannot generate interpersonal norms, and so the public expression of that state plays an essential role, as well. But the norms on the mental state are the original source of the obligation, and they determine what the norm on the publicly expressed mental state is: the norms on the promise derive from the public expression of the speech act of promising, and the norms on the public expression of the speech act of promising derive from (and are explanatorily downstream from) the norms on the mental state conveyed by that speech act.[13] If the norms on this mental state were different, the norms on the speech act would be different, too.[14]

Because *every* promise expresses a resolution—that is, because every promise, regardless of whether the promisor is sincere or believed by the promisee, *conveys* that the promisor has resolved to act—MSF has broader applicability than do the theories discussed in the previous section, and can accommodate the marginal cases that these other theories cannot. When your highly unreliable friend tells you that he is going to pay you back the money you lend him, your expectations about repayment are not raised—but your friend does convey to you that he's resolved to pay you back (even if you doubt that he'll carry out this resolution). You can express resolutions in the absence of social conventions. A promise made to someone who later dies expresses a resolution in just the same way as a promise made to someone who remains living does. And one can convey that she has resolved to Φ without thereby communicating an intention to obligate herself. MSF is broader than the views discussed in the previous section because it derives promissory obligation from a core component of the speech act of promising itself, rather than from the content of particular promises or the various ways in which a promise interacts with the world (e.g., by creating expectations in the promisee, or being part of a social practice involving sanctioning).

[13] The order of explanation goes from mental state to speech act rather than vice-versa, for there is a clear sense in which mental states are independent of and prior to the speech acts that express them, and in which speech acts are not independent of and prior to the mental states they express. One can properly and without any insincerity be in a given mental state without expressing it via a speech act. Since the mental states can properly function entirely independent of the speech acts, it would be strange to derive the norms on them from the norms on speech acts. But speech acts are not independent from mental states in the same way: they express that the agent is in a particular mental state, and it is problematically insincere to have the speech act without that mental state. This dependency makes it quite natural to derive the norms on the speech act from the norms on the mental state.

[14] MSF states that you can obligate yourself at will (pending the promisee's acceptance) by publicly expressing that you have formed a conditional resolution to act. Does this make it a normative power view of the sort discussed in section 1? It doesn't, because all theories of promissory obligation grant that promising involves deliberately creating a new obligation by doing something intentional (e.g., raising someone's expectations, or participating in a social practice, or conveying a certain intention). What makes normative power views distinct is the explanatory structure that they take (i.e., a transcendental argument grounded in the legitimate interest we have in being able to alter the normative situation by fiat). MSF proposes a different explanatory structure: the basis of the promissory obligation is underlying the norms on the mental state conveyed by the promise, and what the public expression of that state commits one to. Thanks to an anonymous referee for discussion of this point.

So far, I have illustrated what MSF can accomplish if it is true. My aim in the rest of this paper is to illustrate how plausible MSF in fact is. I begin by defending the claims that promises express resolutions (claim 1). I defend claims (2) and (3) in subsequent sections.

3 Defending Claim (1): Promises Express Resolutions

What do you express or convey when you perform the communicative act that constitutes making a promise? That is, what mental state do you at least purport to be in when promising, and are you really in if your promise is sincere?[15] A natural idea is that promising to Φ expresses that you plan to Φ, and are serious about carrying out this plan. Someone who promises to Φ communicates that she really will Φ, even if she doesn't feel like doing so at the time, or a better option arises, etc. When I promise you that I will attend your show tomorrow, I'm telling you that I am going to be there, even if faced with barriers to action that might otherwise prevent me from going. If my promise is sincere, I really do have such a plan. If I am insincere, I express to you that I am serious about going to your show without actually being so committed.

How can we cash out this notion of a serious plan? We cannot appeal merely to desires, for desires aren't normatively committing in the way that promises are: promising to Φ pro tanto morally obligates you to Φ, but desiring to Φ clearly does not. Moreover, we often sincerely promise to do things that we don't desire to do; you can sincerely promise your department chair that you'll attend the next faculty meeting, even though this is not how you desire to spend your Friday afternoon.[16] Intentions are a mental state with more stability than desires, as Michael Bratman (1987, 18–20) and others have argued. If I intend to go to your band's show tonight, then I have settled the matter of what I am going to do. I should not continue to deliberate about it or revise my plans for no good reason. Similarly, promising you that I will attend the show settles the matter of what I am going to do. It would be inappropriate to continue deliberating about what to do or to revise my plans

[15] I refer not to sincerity in the sense of being well-meaning or earnest, but to *communicative sincerity*. When A utters a speech act S that expresses state M, A is communicatively sincere if and only if A really is in state M.

[16] You might have other desires that would be satisfied by going to the meeting (e.g., getting on the chair's good side). But we can imagine scenarios in which this is not the case.

doi: 10.48106/dial.v74.i4.03

unilaterally. Intentions might therefore be a decent candidate for the mental state expressed by promising.

However, mere intentions are not stable enough to capture the seriousness of the plan and the strong sense of commitment involved in promising. Suppose that I intend to go to your band's show tonight because I have nothing better to do. You attempt to solicit a promise from me to attend the show. I say, "I promise to be there. And as of now, I intend to go. But you should know that I'm not committed to refraining from revising that intention. My plans might change between now and then, especially if I get a better offer." Such a statement does not seem to be a genuine promise. This is because it is unproblematic for me to intend to go to your band's show tonight and then abandon my intention because a more appealing offer comes along. But such circumstances would not license my breaking a promise to you to see your band play.

This shows us that promises express something stronger and with more stability than typical intentions. Consider the special kind of intentions that we tend to form at the start of a new year—what we often call *resolutions*. These are particularly serious and stable intentions that we're strongly committed to, usually about important goals that we expect might be very difficult to attain. I take resolutions to be intentions that are especially robust or resistant to revision.[17] Resolutions are necessary when you plan to act and care about whether you do so but suspect that some temptation or other barrier to action (such as laziness, aversion to an unpleasant task, apathy, etc.) might cause you to abandon your plan were you not to bolster it somehow. Forming a resolution is one means by which you can bolster your plan and more effectively resist temptation.[18] It is plausible that when I promise to Φ, I express that I have resolved to Φ—that is, that I plan to Φ, and that I care enough about whether I do so that I will not reconsider or abandon that plan, even in the face of temptation, laziness, better offers, and the like. Unlike intentions, it is

[17] The details of how we cash out resolutions do not matter, so long as there is some coherent notion of resolution that implies that we are irrational if we over-hastily revise or fail to act on our resolutions without a good excuse. In Liberman (2016), I argue that resolutions consist in an intention to act coupled with the desire not to reconsider that intention, and offer an objection to Richard Holton's (2009) closely related view according to which a resolution is an intention to act coupled with an intention not to reconsider.

[18] I don't mean to claim that resolutions are the *only* effective means of resisting temptation. There are other means by which you can resist temptation, which can be more effective than resolution-making; the best way to refrain from the temptation to drink tonight might not be to form a resolution, but to lock the liquor cabinet and give the key to a reliable friend.

problematic to abandon a resolution because you no longer feel like acting, or because a better offer comes along. So a resolution-based account will not overgenerate cases of permissible promise-breaking in the way that an intention-based account would.

However, not every publicly expressed resolution counts as a promise; to simply announce in your presence that I have resolved to run a marathon is not to promise you that I will do so. This is because promises require a second party; the acceptance or uptake of the promisee is essential to making a promise. Promissory resolutions must therefore take account of promisee acceptance. They can do this if they are *conditional on* the acceptance of the promisee, in the way my resolution to go for a picnic tomorrow might be conditional on the weather being good. If I offer to promise to run a marathon with you, I convey that I resolve to run *on the condition that you accept* (and do not subsequently reject) my promissory offer. A valid promise is successfully created (and a pro tanto moral obligation generated) only if you accept my offer.

I propose that promises express a resolution to act, conditional on the acceptance of the promisee. However, we might worry that this account overgenerates cases of legitimate promise-making. For we can imagine publicly proclaimed resolutions that are explicitly conditional on someone else's acceptance or agreement, but that do not seem to count as promises. For example, suppose I tell my personal trainer that I resolve to lift weights with her five times a week, but only on the condition that she agree to work with me; she agrees. I have announced that I have resolved to train five days a week, conditional on her acceptance of this resolution, and she has accepted. But I don't seem to have made her a promise. How do we distinguish genuine promises from announcements that one has a conditional resolution?

The response to this worry is simple: it is plausible that a necessary precondition for validly promising is recognizing that you are making a promise in the first place. In general, successfully engaging in an intentional action that alters the normative situation requires a recognition of what you are doing, e.g., you must recognize that you are granting consent in order to successfully do so, and must be aware of the fact that you are transferring property in order to make a gift, etc. Similarly, it is plausible that I need to understand that I am making a promise in order to successfully do so. Likewise, it's plausible that the promisee must be aware of the fact that she is accepting a promise, as well. If the trainer does not *take herself to be accepting a promise* and thereby generating a valid promissory obligation when she says she'll work with me, then what she is doing is agreeing with a resolution, and not accepting a promise.[19]

4 Defending Claim (2): Resolutions Commit You to Acting

4.1 *Resolutions and Rational Commitment*

Someone who resolves to Φ incurs a self-imposed, pro tanto, subjective rational obligation to Φ. I will refer to such obligations as *rational commitments.*[20] The easiest way to get a sense of what I mean by rational commitment is to consider the way in which holding one belief can commit you to holding another. Philosophers frequently talk about the ways in which our beliefs commit us, e.g., because Jack believes that only consequences are relevant for moral assessment, he is committed to believing that he ought to kill one person to save two. Such commitments can come apart from what you objectively ought to believe, all-things-considered: perhaps Jack should not believe that he ought to kill one to save two. A resolution commits you to acting in a

[19] MSF proposes that promises are best understood as conditional resolutions, but we must appeal to the concept of a promise to distinguish which resolutions are promise-generating and which aren't. Is this problematically circular? It isn't, for MSF is not meant to be a descriptive account of what promise-making consists in or a tool for identifying which utterances count as promises and which don't. Rather, MSF is an account of the nature of promissory obligation and of the normative force of promises. MSF posits that this ultimately stems from the norms on the conditional resolution that is publicly expressed when you make a promise. Promisor and promisee must both take themselves to be participating in promise-making in order to determine which resolutions will play this role. But this awareness of what they are doing is not the fundamental normative mechanism and does not provide any deep explanation of how and why promises bind. And in general, it is unproblematic to appeal to the concept of X as one part of an explanation of the nature or normative force of X. For example, suppose I am offering an account of the nature and normative force of giving, according to which A gives X to B (in a moral, rather than a legal sense) if and only if A intends to transfer X to B in an irrevocable way. Some irrevocable transfers will fail to count as gifts—say, those that are made under duress and are perceived by A and B as threats. We can unproblematically state that A must intentionally conceive of themselves as giving X to B in order for X to count as a gift, and that gift-giving generally involves conceiving of oneself as making a gift vs. acceding to a threat. This doesn't diminish the explanatory force of the account of giving as irrevocable transfer; appealing to the *concept* of a gift is part of what cashing out the underlying *nature* and *normative force* of giving requires. Something similar is true for my account of promissory obligation: we must appeal to the concept of a promise to cash out the underlying nature and normative force of promising. Thanks to an anonymous referee for discussion of this point.

[20] For more on the concept of commitments (as distinct from reasons and all-things-considered obligations), see Shpall (2013, 2014), as well as Liberman and Schroeder (2016).

similar way as believing that p commits you to believing the obvious and relevant consequences of p: someone who believes that p and that p entails q but does not believe that q when the question of whether q is salient fails to act on her rational commitments, and as a result her overall set of beliefs is not as complete and coherent as it should be. Similarly, someone who resolves to Φ at a particular time and then fails to intend to do so at that time because she readily abandons this resolution fails to act on her rational commitments, and as a result is not as effective a planning agent as she could be.[21] The rational obligation to act on one's resolutions stems from a broad demand for coherence in one's long-term plans; we are rationally committed to acting on our resolutions because this is essential for effectively carrying out our plans and acting in line with our important goals and values in the long term. To be clear, the obligation to fulfill your resolutions is not the same as a narrow requirement to be instrumentally rational or risk incoherence.[22] Rather, it is grounded in a broader demand for a more holistic sort of coherent planning agency. In order to meet our most important and difficult long-term goals, we must bolster ourselves against succumbing to temptation in ways that would undermine these goals. Resolution-keeping enables us to do this.

We can best illustrate how resolutions incur rational commitments—that is, how they impose subjective, pro tanto, rational obligations—with an example. Suppose you are generally hesitant to try new foods, and are deeply entrenched in the habit of eating pizza for lunch every day. You very much want to expand your culinary horizons, but are such a creature of habit that you are unlikely to do so unless you force yourself into it somehow. So you resolve to go to a Thai restaurant for lunch today, knowing that if you don't form this resolution you are likely to fall back into your pizza habit. When you leave your office to go eat lunch, you abandon your plan to go to the Thai place and head to the pizzeria instead, deciding that you might as well just eat pizza today, since it's easier for you to order from a familiar menu. Something is wrong with you in this picture; resolving to eat Thai food in order to expand your culinary horizons and then changing your mind without good reason is problematically

[21] Time indexing is necessary to avoid over-generating cases: if I resolve to go to eat Thai food at noon, it's no problem that I haven't yet formed the intention to do so at 9 AM. But it would be problematic if lunchtime rolls around and I haven't formed such an intention.

[22] For influential treatments of conditional normative requirements grounded in coherence as the basis of instrumental rationality, see Broome (1999), who articulates a wide-scope view, and Kolodny (2005), who articulates a narrow-scope view; see also Way (2010) for an overview of this debate.

doi: 10.48106/dial.v74.i4.03

self-undermining. It is irrational or incoherent to resolve to eat Thai food because you care about broadening your horizons, and then decide to stick with your pizza habit because it is easier. It's not that eating pizza every day is *independently* irrational; having pizza all the time is rationally permissible, if boring. Rather, it's that it is irrational to *resolve* to eat Thai food, and then abandon this resolution for no good reason.

Compare abandoning a resolution to eat Thai food to merely *desiring* to have Thai food and eating pizza instead. This is perfectly acceptable; failing to act on a particular desire is not irrational. Or compare it to the case in which you *intend* to have Thai food but don't really care about whether this plan changes. You do not display any irrationality if you change your mind because you suddenly have a craving for a sandwich; you've simply made a permissible change of intention on the basis of a change in desire. Changing your resolution on a similar basis is not so innocuous. When you abandon your resolution to go to the Thai restaurant for lunch, you are undermining your own endorsed goals and plans: you value culinary diversity, and adopt this as one of your goals, but do not succeed in attaining it.

We can best see how abandoning a resolution without good reason is irrational by comparing a pair of similar cases. Suppose that you and I each have a reason of strength X to eat Thai food and expand our respective culinary horizons. And suppose we each care about expanding our culinary horizons to the same extent. But only I take action about it: I *resolve* to eat Thai food, making it a part of my plan and adopting it as one of the concerns that I will focus on. If I change my mind and have pizza for lunch, I go wrong in a way that you do not when you have pizza. This is not to say that it is *never* permissible to abandon a resolution. Sometimes, there are weightier considerations in favor of revising a resolution than there are in favor of maintaining it—say, if you resolve to have Thai food and an old friend unexpectedly offers to meet you for lunch at the pizza place, or you realize that Thai food often contains ingredients to which you are allergic. However, resolutions rationally commit you to acting unless there is a good excuse in place.

4.2 *Resolutions and Moral Obligation*

It might seem prima facie puzzling to claim that the norms on promises can be explained by appeal to the norms on resolutions. I've argued in the previous section that resolving *rationally* commits you to acting. But promising *morally* obligates you to act, as well. Promissory obligations are also directed towards another person in a way that most resolutions are not: a promise does not create a generic moral obligation to act in a certain way, but is owed *to the promisee,* who is uniquely wronged should the promise be broken. We can assess the directedness of an obligation by asking whether anyone is uniquely positioned to forgive (or has special standing to resent) the violator, where resentment is understood as a negative evaluative attitude that is appropriate only in response to a personal violation. For example, suppose a passenger on an airplane has a medical emergency. The flight crew asks if any medical professionals are on board. There is a doctor on board (of medicine, not of philosophy!) who could easily assist the passenger, but who fails to do so because she prefers not to miss any scene of her in-flight movie. The other passengers on the plane might criticize or blame the doctor for her callous behavior. But the distressed passenger has unique standing to resent or forgive the doctor in a distinctive way. The doctor owes it *to the distressed passenger* to help him; this is not a generic obligation to be beneficent, but an obligation owed to a particular person. How can a resolution yield a commitment that is both *moral* and *directed* in this sort of way?

I have three responses to this important question. First, I want to ease the explanatory burden on MSF. I am not trying to offer a complete or exhaustive account of all of the ways in which promises can morally obligate us, which means that MSF does not need to explain all of our intuitions about the robust moral force of promises. Rather, there are often *multiple* sources of moral obligation to keep any given promise, and these all contribute to the overall moral force of the promise.[23] These sources are the morally important considerations raised by the other theories of promissory obligation discussed in section 1: I ought to keep my promise to you to pick you up from the airport because it will harm you if I don't, *and* because failing to do so would problematically free-ride on a valuable social practice, *and* because I have publicly conveyed to you that I have resolved to do so. This last source of obligation stems from MSF, and is present in all cases of promising.

Consider a pair of examples to illustrate how the moral force of a promise can vary (and will be stronger when other considerations are in play). First, suppose you promise your mother that you will not sell a treasured family heirloom that is in your possession. Assume that there aren't any strong reasons for you to sell it; you're doing fine financially, and the heirloom is

[23] When there are multiple sources of obligation to keep a promise, the obligation is over-determined, and the moral force of it is stronger; see the case about reasons not to sell a family heirloom below.

doi: 10.48106/dial.v74.i4.03

not worth very much. The moral reasons you have to keep this promise are not exhausted by the fact that you've conveyed to your mother a resolution to hang on to the heirloom; they also stem from harm avoidance, family loyalty, the importance of maintaining family traditions, and the like. And so the moral force of the promise to refrain from selling the heirloom will be quite strong.

In cases in which no other reasons to keep a promise are present—that is, in which no expectation is formed, in which no trust can be violated, etc.—we see that promises generate comparatively weak moral obligations. Second, suppose that we are airplane seatmates stranded on the tarmac because of a flight delay. I ask you to promise me that you will write a negative review about the airline's customer service when you get off the flight, and you make the promise (and thereby resolve to do so, conditional on my acceptance). Assume that there are no independent sources of moral obligation to keep this promise—we are strangers, I will never know whether you write the review, doing so will in no way affect the airline's business, etc. You nevertheless seem to be pro tanto morally obligated to write the review. If you have an opportunity to easily write a negative review and you fail to do so without a good excuse, you will not be entirely as you ought to be. But neither will you be failing morally in a drastic way, for your obligation to keep the promise to me is not an especially important one. MSF purports only to explain this type of obligation: the relatively weak moral force that stems from the bare act of promise-making itself, including in marginal cases. This mitigates the explanatory demands on MSF, which needs to explain only this form of relatively weak moral obligation.

Second, the conditionality of the resolution on the interest of the promisee is what enables it to yield a directed obligation. Successful promise-making requires the acceptance of the promise by the promisee, and I've argued that promissory resolutions are conditional on such acceptance. Accepting a promise implies that you are interested in the promise being made, and (usually) that you are interested in it being kept.[24] It's plausible that resolutions can generally yield directed obligations when the interests of another person

[24] In typical cases, the promisee accepts the promise because she wants the promised action to occur. In deviant cases, the promisee might accept the promise for other reasons; Vera Peetz (1977) describes a case in which you accept your neighbor's promise to give you homemade jam not because you actually want her jam, but because you do not want to hurt her feelings by declining her offer. In this case, you are interested in the promise's being made—for this is necessary to spare your neighbor's feelings—even though you do not care whether it is kept.

are directly involved in the core content of the resolution. Another person's interest in how you are going to act is normatively relevant, and in some circumstances—including promissory resolutions—this can ground a directed obligation.

We can illustrate how another person's interest in a resolution can yield a directed obligation by considering a series of cases. Suppose you resolve in secret to mow your elderly neighbor's lawn as a favor to him. In doing so, you incur a rational commitment. If you fail to keep the resolution without a good excuse for changing your mind, you will have acted irrationally, but will not have done anything morally problematic. If you inform your neighbor that you have resolved to mow his lawn, you will also incur a directed, expectation-based obligation to your neighbor to either mow the lawn or alter his expectations. And if you neither mow the lawn nor warn him in advance that you do not plan to do so, your neighbor is entitled to resent or forgive you for violating this expectation-based obligation. So too would be any third party who has a stake in the matter and whose expectations were similarly raised by your announcement of your resolution; if your neighbor's landscaper was present when you announced your resolution, she is likewise entitled to resent you for neither mowing the lawn nor warning her that you've changed your mind.

However, suppose you inform your neighbor that you have formed a resolution that is explicitly conditional on his interests—you say, "I've resolved to mow your lawn this weekend, so long as you want me to." In this case, it is somewhat plausible that you incur a directed obligation to your neighbor to mow the lawn. For your resolution conveys that you plan to mow his lawn, and that this plan depends not on *your* interests or desires, but on *his*. The conditionality of this resolution places the plan in his hands. If you don't mow the lawn and instead simply warn your neighbor ahead of time that you have changed your mind, you have perhaps done something morally problematic. It would not seem terribly out of place for your neighbor to resent or forgive you, in a way that would be odd if the resolution were not conditional on his interests. But if you warn the landscaper ahead of time that you will not be carrying out the conditional resolution as your neighbor wants you to, it *would* seem out of place for the landscaper to resent or forgive you; you have discharged your expectation-based obligation to her, and owe nothing more.[25]

[25] There might be cases in which it is morally inappropriate for independent reasons to give the landscaper a warning instead of mowing the lawn yourself—say, if you know that the landscaper

doi: 10.48106/dial.v74.i4.03

It is even more plausible that a directed obligation is formed if your resolution is explicitly conditional not on your neighbor's *desire* that you mow his lawn, but on his active *acceptance* of your resolution to do so—that is, if you say, "I've resolved to mow your lawn, but will only do so if you remain actively on board with this plan." For in that case, you have conveyed to your neighbor that you plan to mow his lawn, and that this plan depends not on your whims or desires but on his active endorsement of the plan. If you change your mind and don't mow the lawn in spite of his continued uptake, it seems appropriate for him to resent or forgive you, even if you do warn him in time. Again, the landscaper does not seem similarly positioned to resent or forgive you, since the resolution was not conditional on her agreement. Your resolution yields not just a generic obligation owed to anyone who overheard you, but an obligation that is directed specifically towards your neighbor. In general, if you resolve to Φ conditional on the agreement, acceptance, or uptake of A, it seems that you owe a directed obligation to A to Φ. Promissory resolutions are always directed in this sort of way, because they are always conditional on the acceptance of the promisee.

Third, conditional resolutions generate moral obligations because resolving conditionally on A's acceptance and then failing to act when this condition is met fails to take proper consideration of A's interests. To fail to adequately account for another person's interests when you are engaged in a direct interaction with them is an interpersonal, moral sort of failing. It betrays the wrong kind of attitude to take towards another person, and can be construed as a form of disrespect or a problematic lack of moral concern. For example, suppose I proclaim to my family that I've resolved to give a particular heirloom to my cousin A, so long as she agrees to take it. She agrees, but I change my mind and give it to a different cousin, B. Assume A has no independent claim over the heirloom that B lacks; had I not formed a conditional resolution, it would be permissible for me to give the heirloom to A or B. Since I expressed a conditional resolution to give the heirloom to A, I seem to be *slighting* or *wronging* A when I give the heirloom to B instead. For I am failing to take her interests into account as I should and to give them their proper weight. My expressing a resolution conditional on A's interests entails that she must be given special consideration; in the case where I'm simply deciding between A and B without expressing any resolution, I don't wrong or slight A by choosing

has rearranged her entire schedule because she thought she didn't need to mow your neighbor's lawn. But generally, expectation-based obligations can be satisfied either by acting or altering the expectations.

B, even if A wants the heirloom more than B does. Promissory resolutions are those in which you are morally required to take special consideration of the promisee's interests. To fail to do so is to fail to properly respect the promisee, just as failing to give the heirloom to A when she has accepted my conditional resolution fails to properly respect her. Granted, this may not be a very serious or significant moral failing. But as we saw above, MSF is burdened with establishing the existence only of a relatively weak sort of moral obligation.

In the last two sections, I have offered arguments for two of the core claims that the argument for MSF relies on: (1) that promises express resolutions; (2) that resolutions rationally and morally commit you to acting, all else being equal. These claims are interesting in themselves, as they help us understand the relationship between promises and resolutions, as well as the nature of the commitments incurred by resolutions. With the addition of a third claim—that is, the determination claim—this relationship between promises and resolutions becomes significantly more interesting, for it can generate a new kind of theory of promissory obligation.

At present, I cannot fully defend the determination claim or offer a deep explanation of why it holds. However, I can offer evidence to suggest that the determination claim is both plausible and theoretically fruitful, insofar is it can help explain the norms on speech acts other than promising.

5 Defending Claim (3): Evidence for the Determination Claim

The third claim on which MSF depends is the determination claim, which states that the norms governing the mental state expressed by a speech act (at least partially) determine what the norms on that speech act are, in both sincere and insincere cases. In other words, the determination claim tells us that saying that you're in state X commits you to behaving as if you are in that state, regardless of whether you in fact are.

My first piece of evidence for the determination claim is the observation that, in general, we need to assume the truth of something like the determination claim in order to have fair and productive social interactions. The determination claim states that you must act as if you really are in a particular mental state when you convey to others that you are in that state. People are not mind-readers, and our conversations and social interactions are generally

doi: 10.48106/dial.v74.i4.03

presumed to be cooperative. In light of this, it would be unfair to expect people to be able to ascertain when our utterances are sincere and when they are not. It follows that we must be able to take what people say at face value if we are to have productive interactions with them, at least in typical circumstances.[26] Doing so enables us to respond appropriately to them: to predict what they might do and say next, and to alter our behavior in light of theirs, etc.

If we weren't entitled to presume that people really were in the states they purport to be in, we wouldn't be able to interact with them very effectively. If I promise to Φ, my promisee must be entitled to presume that I am committed to Φing, lest she be at a loss for how to respond to me. And because she cannot know my inner mental state, it would be unreasonable and unfair to expect anything else of her. Accordingly, my promisee is entitled to interact with me as if I have in fact resolved to Φ—which might involve her believing that I plan to Φ, structuring her future plans and behavior around my Φing, or simply responding appropriately in the moment to my commitment to Φing. And as promisor, I should in turn behave as if I really have resolved, and thereby enable my interlocutor to take my promissory utterance at face value—which is to say, I should act in accordance with the determination claim.

The rest of my evidence for the determination claim is circumstantial: there exist other cases in which publicly conveying that you are in a mental state M by performing a speech act S commits you to acting in whatever way is required by M, regardless of whether you really are in M. Promising is not unique in this regard, but is part of a general pattern. This gives us good reason to think that the determination claim is broadly true, and is not merely an *ad hoc* principle that applies only to the case of promises and that I am invoking out of the blue to defend MSF.

[26] Because the mental state is explanatorily prior to the speech act (see footnote 13), the mental state is also explanatorily prior to the action that results from the speech act. Suppose that I am in state X, and I convey this by saying "I'm in X." How others can reasonably expect me to act on the basis of this utterance is fundamentally determined not by my utterance that I'm in X, but by what *being in state X* commits me to. To be fair to others, I must act as if I am in state X, unless there is some reason not to take my utterance at face value—say, if it is mutually understood that I am in a strategic (e.g., game-playing) context.

5.1 *Forgiving*

It is plausible that forgiving someone expresses that you have repudiated or foresworn blaming them.[27] Repudiating blame of A for doing X plausibly commits one to ceasing blaming A for X in the moment, and to refraining from actively expressing blame towards A for X again in the future. Publicly expressing forgiveness—regardless of whether you have privately foresworn blame—likewise seems to commit you to refraining from expressing blame in these same ways. That is, publicly conveying that you have repudiated blame by performing a speech act of forgiveness commits you to behaving as if you really have repudiated blame.

To illustrate, suppose that Anna forgets that today is her wedding anniversary, and she fails to meet her spouse Betty for a celebratory dinner they have planned. Betty knows that she is likely to hold this mistake against Anna, and doesn't want to damage their relationship by doing so. So she decides to foreswear blaming Anna for her oversight. This private foreswearing of blame commits her to refraining from expressing blame towards Anna. And telling Anna that she has forgiven her likewise commits Betty to refraining from expressing blame towards Anna. This is so even if Betty is insincere, and forgives Anna not because she has in fact foresworn blaming her, but because she wants to avoid conflict. Betty's interpersonal utterance of forgiveness nevertheless commits her to refraining from openly expressing blame. As with promising, the public expression of a foreswearing of blame changes the nature of the norms to which you are subject; it transforms a private commitment to refraining from engaging in blaming activities into an interpersonal, directed obligation to avoid continuing to blame the wrongdoer in the future. Finally, and again as with promising, this is compatible with the existence of additional explanations of why you should refrain from expressing blame in particular cases; perhaps Betty has an obligation to refrain from expressing blame because she has publicly expressed that she has forgiven Alice, *and* because she is independently obligated to avoid acting unfairly, and expressing blame would be unfair since Betty forgot their anniversary last year.

[27] This is a common view of forgiveness in both everyday practice and the philosophical literature; the view is usually attributed first to Bishop Butler. See discussion of this view in Griswold (2007).

5.2 *Apologizing*

Similarly, apologizing for Φing expresses that you regret or are sorry that you Φed. When someone regrets an action in this way, they incur a commitment to actively taking responsibility for it somehow. What this involves will vary in different cases; a cheating spouse who regrets their infidelity is obligated to avoid straying again, while a party guest who regrets spilling red wine on a white rug incurs a commitment to clean up the spill. Philosophical and popular consensus is that someone who publicly apologizes for Φing likewise commits themselves to taking responsibility for their action.[28] And this is so regardless of whether the apology was sincere; publicly conveying that you regret your action by apologizing commits you to behaving as if you really have regretted your action.

For example, suppose Christa catches her student Danny using his phone during class, in violation of her policy. Danny apologizes to Christa, which commits him to taking responsibility for his error (e.g., by admitting that he was wrong and refraining from using his phone in class again). This is so even if Danny is insincere, and apologizes only because he fears that Christa will dock his participation grade if he does not. As with promising and forgiveness, a public apology transforms a personal feeling of regret that privately commits you to making amends into an interpersonal demand to make such amends. There may also be additional moral considerations present that require agents to take active responsibility for their wrongdoing. But even if these considerations are not present, the mere (sincere or insincere) expression of regret commits you—at least in a weak way—to taking responsibility for your wrongdoing.

5.3 *Asserting*

There is much disagreement about what mental state assertion expresses and whether there are norms of further commitment on assertions; adjudicating between these views is too large a project to be adequately handled here. But we can assume a particular view to illustrate how assertion might pattern

[28] For example, Mihaela Mihai (2013) notes that while philosophical accounts of apology vary, "there is a growing consensus that an authentic apology implies an acknowledgement that the incident in question did in fact occur and that it was inappropriate, a recognition of responsibility for the act, the expression of an attitude of regret and a feeling of remorse, and the declaration of an intention to refrain from similar acts in the future."

with promising, forgiving, and apologizing. Suppose for the sake of argument that assertion expresses belief (which is compatible with assertion expressing something else that includes belief as a component, such as justified belief or knowledge). If you believe that *p* and are questioned about whether *p*, it is plausible that you are normatively committed to defending or justifying *p*. It is also plausible that this same commitment is inherited by assertions: some philosophers argue that someone who asserts that *p* takes on a special commitment to the truth of *p*, which can be cashed out as a commitment to justify or defend *p* to one's interlocutors.[29] This is plausibly so even in insincere cases; if you are not going to retract your insincere assertion that *p*, you should be prepared to defend or justify it. If this is the case, then assertion patterns with the other speech act/mental state pairs we've been discussing: publicly conveying that you believe *p* by asserting that *p* commits you to behaving as if you really believe *p*, and transforms your private commitment to behave as if *p* is true into a commitment to defending *p* publicly.

5.4 *Conclusion*

The determination claim is a natural and plausible explanation of the pattern outlined above, for two reasons. First, speech acts are dependent on the mental states they express, in a way that mental states are not dependent on the speech acts used to express them. You can properly and without any insincerity be in a particular mental state without expressing it via a speech act; it is no problem to resolve to act without promising that you will do so. But all of the speech acts we have been discussing express that the agent is in a particular mental state, and it is communicatively insincere to perform the speech act without being in that mental state. The direction of explanation proposed by the determination claim tracks this dependency; it would not make sense for the direction of explanation to go in the other direction.

Because we have reason to think the determination claim is true—and we also have reason to think that promises express resolutions, and that resolutions rationally (and sometimes morally) commit us to acting—we have reason to think that MSF is a viable theory of promissory obligation. And MSF has appealing explanatory advantages. Since the mental state expressed by a promise is both distinctive of promises and present in all cases, MSF

[29] See, among others, Peirce (1935), Searle (1969), Brandom (1983), Wright (1992), Watson (2004), and MacFarlane (2005).

captures the minimal, essentially promissory obligation that is always there, while being open to the pluralist idea that other theories can explain why we have stronger moral reasons to keep our promises in many cases. Moreover, the success of MSF should lead us to be optimistic about the possibility of providing similarly structured and equally resourceful accounts cashing out the norms on other sorts of speech acts in terms of their underlying mental states.*

Alida Liberman
0000-0002-5182-569X
Southern Methodist University
aliberman@smu.edu

References

Brandom, Robert B. 1983. “Asserting.” *Noûs* 17(4): 637–650, doi:10.2307/2215086.

Bratman, Michael E. 1987. *Intentions, Plans and Practical Reason*. Cambridge, Massachusetts: Harvard University Press.

Broome, John A. 1999. “Normative Requirements.” *Ratio* 12(4): 398–419, doi:10.1111/1467-9329.00101.

Cholbi, Michael J. 2002. “A Contractualist Account of Promising.” *The Southern Journal of Philosophy* 40(4): 475–491, doi:10.1111/j.2041-6962.2002.tb01913.x.

Cottingham, John G. 1985. “Review of Robins (1984).” *The Philosophical Quarterly* 35(140): 315–318, doi:10.2307/2218913.

Downie, R. S. 1985. “Three Accounts of Promising.” *The Philosophical Quarterly* 35(140): 259–271, doi:10.2307/2218905.

Griswold, Charles L. 2007. *Forgiveness. A Philosophical Exploration*. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511619168.

Heuer, Ulrike. 2012. “Promising – Part 1.” *Philosophy Compass* 7(12): 832–841, doi:10.1111/j.1747-9991.2012.00524.x.

Holton, Richard. 2009. *Willing, Wanting, Waiting*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199214570.001.0001.

* This paper is adapted from my dissertation, *The Mental States Theory of Promising* (2015), completed at the University of Southern California. More people than I can list here have helped me to articulate and refine these ideas over the years. Special thanks are due to my committee members Steve Finlay, Jake Ross, and Gary Watson, to my graduate school classmates (especially Josh Crabill, N.G. Laskowski, Michael Milona, Caleb Perl, and Abelard Podgorski), to audiences at USC and Bowling Green State University, and most importantly to my dissertation supervisor Mark Schroeder.

Hooker, Brad. 2011. “Promises and Rule-Consequentialism.” in *Promises and Agreements: Philosophical Essays*, edited by Hanoch Sheinman, pp. 235–252. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780195377958.001.0001.

Hume, David. 1739. *A Treatise of Human Nature*. London: John Noon, at the White-Hart.

Kolodny, Niko. 2005. “Why Be Rational?” *Mind* 114(455): 509–563, doi:10.1093/mind/fzi509.

Kolodny, Niko and Wallace, Richard Jay. 2003. “Promises and Practices Revisited.” *Philosophy & Public Affairs* 31(2): 119–154, doi:10.1111/j.1088-4963.2003.00119.x.

Lemos, Noah M. 1987. “Review of Robins (1984).” *Philosophy and Phenomenological Research* 47(4): 685–688, doi:10.2307/2107242.

Liberman, Alida. 2015. “The Mental States First Theory of Promising.” PhD dissertation, Los Angeles, California: Philosophy Department, University of Southern California, http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll3/id/591683.

—. 2016. “Reconsidering Resolutions.” *Journal of Ethics and Social Philosophy* 10(2), doi:10.26556/jesp.v10i2.98.

—. 2020. “Consequentialism and Promises.” in *The Oxford Handbook of Consequentialism*, edited by Douglas W. Portmore, pp. 289–309. Oxford Handbooks. Oxford: Oxford University Press, doi:10.1093/oxfordhb/9780190905323.013.16.

Liberman, Alida and Schroeder, Mark. 2016. “Commitment: Worth the Weight.” in *Weighing Reasons*, edited by Errol Lord and Barry Maguire, pp. 104–120. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199315192.003.0005.

MacFarlane, John. 2005. “Making Sense of Relative Truth.” *Proceedings of the Aristotelian Society* 105: 305–323, doi:10.1111/j.0066-7373.2004.00116.x.

Marušić, Berislav. 2013. “Promising Against the Evidence.” *Ethics* 123(2): 292–317, doi:10.1086/668704.

Mihai, Mihaela. 2013. “Apology.” in *Internet Encyclopedia of Philosophy*. University of Tennessee at Martin, https://iep.utm.edu/apology/.

Owens, David. 2012. *Shaping the Normative Landscape*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199691500.001.0001.

Peetz, Vera. 1977. “Promises and Threats.” *Mind* 86(344): 578–581, doi:10.1093/mind/LXXXVI.344.578.

Peirce, Charles Sanders. 1935. *Collected Papers of Charles Sanders Peirce, vol. 5: Pragmatism and Pragmaticism*. Cambridge, Massachusetts: Harvard University Press. Edited by Charles Hartshorne and Paul Weiss.

Pink, Thomas. 2009. “Promising and Obligation.” in *Philosophical Perspectives 23: Ethics*, edited by John Hawthorne, pp. 389–420. Hoboken, New Jersey: John Wiley; Sons, Inc., doi:10.1111/j.1520-8583.2009.00177.x.

RAWLS, John. 1971. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press. Revised edition: Rawls (1999).

—. 1999. *A Theory of Justice*. Cambridge, Massachusetts: Harvard University Press.

RAZ, Joseph. 1977. "Promises and Obligations." in *Law, Morality and Society. Essays in Honour of H.L.A. Hart*, edited by Peter M. S. HACKER and Joseph RAZ, pp. 210–228. Oxford: Oxford University Press.

ROBINS, Michael H. 1984. *Promising, Intending and Moral Autonomy*. Cambridge: Cambridge University Press.

SCANLON, Thomas Michael. 1990. "Promises and Practices." *Philosophy & Public Affairs* 19(3): 199–226.

—. 1998. *What We Owe to Each Other*. Cambridge, Massachusetts: The Belknap Press of Harvard University Press.

SEARLE, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

SHEINMAN, Hanoch. 2008. "Promise as Practice Reason." *Acta Analytica* 23(4): 287–318, doi:10.1007/s12136-008-0033-1.

SHIFFRIN, Seana Valentine. 2008. "Promising, Intimate Relationships, and Conventionalism." *The Philosophical Review* 117(4): 481–524, doi:10.1215/00318108-2008-014.

SHPALL, Samuel. 2013. "Wide and Narrow Scope." *Philosophical Studies* 163(3): 717–736, doi:10.1007/s11098-011-9841-z.

—. 2014. "Moral and Rational Commitment." *Philosophy and Phenomenological Research* 88(1): 146–172, doi:10.1111/j.1933-1592.2012.00618.x.

SMITH, Holly M. 1987. "Review of Robins (1984)." *Noûs* 21(4): 604–608, doi:10.2307/2215676.

SOUTHWOOD, Nicholas and FRIEDRICH, Daniel. 2009. "Promises Beyond Assurance." *Philosophical Studies* 144(2): 261–280, doi:10.1007/s11098-008-9209-1.

THOMSON, Judith Jarvis. 1990. *The Realm of Rights*. Cambridge, Massachusetts: Harvard University Press.

WATSON, Gary. 2004. "Asserting and Promising." *Philosophical Studies* 117(1–2): 57–77, doi:10.1023/B:PHIL.0000014525.93335.9e.

WAY, Jonathan. 2010. "The Normativity of Rationality." *Philosophy Compass* 5(12): 1057–1068, doi:10.1111/j.1747-9991.2010.00357.x.

WRIGHT, Crispin. 1992. *Truth and Objectivity*. Cambridge, Massachusetts: Harvard University Press.

A Puzzle About Parsimony

Peter Finocchiaro

In this paper, I argue for the instability of an increasingly popular position about how metaphysicians ought to regard parsimony. This instability is rooted in an unrecognized tension between two claims. First, we as metaphysicians ought to minimize the number of ontological kinds we posit. Second, it is not the case that we ought to minimize the number of ideological expressions we employ, especially when those expressions are of the same ideological kind (e.g., the compositional predicates 'is a part of' and 'overlaps'). I argue that the two claims are in tension with one other. At the very least, minimizing the number of ontological kinds posited entails minimizing the number of expressions employed—more specifically, the "ontologically committing" predicates. But, plausibly, the tension runs deeper than that. I suggest that minimizing the number of ontological kinds just is a specific way of minimizing the number of ideological expressions employed in stating a theory. The two activities target the same aspect of reality, the world's metaphysical structure. I end by evaluating three different responses to this puzzle. Ultimately, I suggest that metaphysicians should treat the minimization of the number of ideological expressions as more important than it currently is treated.

Parsimony is among the most prominent methodological considerations in metaphysics. Yet beneath the surface there lurks a puzzle. I will bring this puzzle about parsimony to light. As I will show, the puzzle highlights a conceptual tension between several prominent positions in metaphysics. I will then offer three responses to the puzzle. Each response faces unique challenges.

First, I will make some starting assumptions. These assumptions are not unassailable. But each is independently plausible and each has broad support amongst metaphysicians.

Parsimony has traditionally been restricted to ontology: do not multiply entities beyond necessity. Lately, however, metaphysicians have turned their attention toward *ideological parsimony*. Ideological parsimony, as I understand it, concerns the primitive (i.e., undefined) terminology used to state a theory. Recently, many philosophers (Brenner 2017; Cowling 2013; Schaffer 2015;

Sider 2011; Turner 2015) have defended the claim that both ontological and ideological parsimony make a theory more worthy of our endorsement. I will assume that they are right.

I will also assume what is sometimes called a *realist* or *externalist* interpretation of ideology. Like an analogous interpretation of ontology, ideological externalism states that the quality of a theory's ideology is ultimately judged by the extent to which it corresponds to objective reality, i.e., the world's metaphysical structure.[1] (Ideological externalism can be contrasted with ideological internalism, which states that the quality of a theory's ideology is judged by details internal to the theoretic process—e.g., the intelligibility of the terminology employed.)

I will also adopt the orthodox approach to meta-ontology, *neo-Quineanism*. According to neo-Quineanism, a theory's ontological commitments are determined by what the theory quantifies over when regimented with a suitably perspicuous language.[2] Finally, I will focus on theories about the fundamental nature of the world. While there might be versions of this puzzle that extend to non-fundamental theories, I do not have much to say about them. That is in large part because I do not have much to say in general about the relationship between fundamental and non-fundamental theories.

These assumptions help generate a puzzle, one that highlights a conceptual tension in how some metaphysicians understand the role of parsimony in theory choice. This tension has, until now, gone unrecognized. To bring out the tension, I identify in section 1 four approaches to parsimony that differ along two axes: ontology/ideology and quantitative/qualitative. We seem to have an intuitive grasp on these approaches and understand the differences between them, in particular the differences between ontological and ideological parsimony. But in section 2, I argue that qualitative ontological parsimony entails a restricted version of quantitative ideological parsimony. This is a surprising and worrisome puzzle. It is surprising because it goes against our intuitive grasp of parsimony. It is worrisome because it seems inconsistent with a popular position amongst metaphysicians—i.e., that greater qualitative ontological parsimony makes a theory more worthy of endorsement but it is not the case that greater quantitative ideological parsimony makes a theory more worthy of endorsement. I then suggest that the entailment is no coincidence; qualitative ontological parsimony may be conceptually distinct

[1] I discuss ideological externalism (as well as ideological internalism) in greater depth in Finocchiaro (2021, 963–969). See, also, Cowling (2013, 3983) and Sider (2011, 13).

[2] See, inter alia, Quine (1948); van Inwagen (1998); Lewis and Lewis (1970).

from quantitative ideological parsimony, but the most sensible applications of them target the same feature of reality, the world's metaphysical structure.

In section 3, I discuss three available responses to this puzzle. First, we could resist the puzzle by rejecting neo-Quineanism. Second, we could downplay the significance of the puzzle by offering a more nuanced application of parsimony. Finally, we could reevaluate the value of quantitative ideological parsimony as a theoretical virtue. Ultimately, I favor the third response. Metaphysicians should value quantitative ideological parsimony more than they currently do.

1 Four Different Approaches to Parsimony

Many metaphysicians think that parsimony should play a role in theory choice. They have cited parsimony in support of theories as wide-ranging as compositional nihilism (Horgan and Potrč 2008), bundle theory (Paul 2017), materialism (Churchland 1984), and nominalism (Melia 2008).

But such metaphysicians often differ in how they use parsimony. Even when restricted to the ontology of a fundamental theory, there are two importantly different approaches they take. Some (e.g., Nolan 1997) tend to prefer the theory that minimizes the number of entities posited. Others (e.g., Lewis 1973) tend to prefer the theory that minimizes the number of *kinds* of entities. Following a convenient shorthand from Cowling (2013), I will name these two different approaches (NO-Parsimony) and (KO-Parsimony), respectively.

I won't take a stand on which approach is best.[3] I simply note that even those inclined toward (NO-Parsimony) also tend to be inclined toward (KO-Parsimony). More generally, among the metaphysicians who care about parsimony at all, most of them accept (KO-Parsimony).[4]

We can also consider the parsimony of a fundamental theory's ideology. David Lewis, for example, claims that modal realism enables us "to reduce the diversity of notions we must accept as primitive" (1986, 4). Theodore Sider argues that compositional nihilism "allows us to eliminate the extra-logical (or perhaps quasi-logical) notion of 'part' from our ideology" (2013, 239). Both modal realism and compositional nihilism are ideologically parsimonious.

[3] For two defenses of different uses, see Lewis (1973) and Tallant (2013).

[4] For instance, Nolan (1997, 330) says "I claim that not only ought we not multiply types of entities beyond necessity, but that we should also be concerned not to multiply the entities of each type more than is necessary."

doi: 10.48106/dial.v74.i4.04

For Lewis and Sider, the ideological parsimony of their theories provides a reason to endorse them.

Just as with ontology, there are two importantly different approaches to ideological parsimony. Metaphysicians may prefer the theory that minimizes the total number of terms that are employed but undefined within the theory ("bits of ideology"). Or they may prefer the theory that minimizes the number of kinds of terms so employed ("ideological kinds"). Adopting another shorthand from Cowling (2013), I will name these approaches (NI-Parsimony) and (KI-Parsimony), respectively.[5]

I should note that it's not obvious how to individuate ideological kinds. (The same could be said about ontological kinds.) Metaphysicians often rely on the imprecise but intuitive method of individuation by topic. For instance, there is an ideological kind corresponding to color. All color predicates like 'blue,' 'periwinkle,' and 'Pantone 19-4052' are of this kind, as are relational predicates like 'is more saturated than.' There is also an ideological kind corresponding to modality. Primitive modal operators, predicates like 'possibly true' and 'consistent,' as well as primitive dispositional predicates like 'fragile' are of this kind. There is much more worth saying about the individuation of ideological kinds.[6] Yet I do not think that my main argument is affected by this issue. In what follows I will stick to the intuitive understanding just sketched.

Some metaphysicians may deny that our use of ideological parsimony can be neatly divided into (NI-Parsimony) and (KI-Parsimony). Yet the distinction seems intuitive enough and many think there is something to it (e.g., Cameron 2012, 18; Cowling 2013, 3897). In addition, there are intuitive reasons to favor (KI-Parsimony) and reject (NI-Parsimony). For one, (NI-Parsimony) seems to force us to make objectionably arbitrary decisions. (NI-Parsimony) recommends that, all else being equal, we minimize the number of compositional predicates in our ideology. What this recommendation precisely amounts to will depend on the resolution of issues that are too large to address here.[7] To see the worry, though, suppose that there are no other relevant considerations regarding our choice of compositional ideology. (NI-Parsimony) then recommends that we employ a minimal expressively adequate set of predicates. For composition, this can be achieved by choosing one from among 'is

[5] Some characterize ideology as concerning the *concepts* employed in stating a theory. I prefer my *linguistic* characterization, for reasons I state in Finocchiaro (2021, 961–963).

[6] I do say much more in Finocchiaro (2019a). See, also, Cowling (2013) and Lewis (1986).

[7] For example, it depends in part on whether composition is classically extensional true (see Parsons 2014, 4).

a part of,' 'is a proper part of,' and 'overlaps' (supplemented with identity). We are then faced with an unsettling question: which of these three should we choose? Each option is unsavory because they all seem to commit us to an unreasonable view about the fundamental compositional structure of the world. Each option also seems impossible to motivate—what could justify choosing one over the other? These worries about arbitrariness disappear if we reject (NI-Parsimony) in favor of (KI-Parsimony). Compositional predicates are (plausibly) of the same ideological kind. So there is no methodological pressure to arbitrarily choose one predicate over the others.[8]

Many metaphysicians nowadays think that both ontological parsimony and ideological parsimony should play a role in theory choice. Why? Historically, parsimony-based considerations have been defended on non-alethic grounds: an ideologically parsimonious theory might be easier to comprehend, or an ontologically parsimonious theory might be more aesthetically pleasing. But such defenses are less popular nowadays since they are seen as relying on reasons that should be irrelevant to theory choice in metaphysics. Nowadays, most metaphysicians who think that parsimony should play a role in theory choice think so because they think parsimony is truth-conducive.[9] This connection between parsimony and truth holds for both ontological parsimony and ideological parsimony. According to ideological externalism, a more ideologically parsimonious theory conveys a more simple—and therefore more likely to be true—picture of the world's structure. Yet metaphysicians are less willing to extend this defense to quantitative ideological parsimony. (Intuitively, a theory that employs only 'is a part of' is not any more likely to be true than a theory that employs 'is a part of' and 'overlaps.') Thus, that approach to parsimony is under-motivated. Because of this lack of motivation and the aforementioned worries about arbitrariness, many metaphysicians reject (NI-Parsimony).

Thus far, I have presented four approaches to parsimony. I have suggested that the overall most attractive package for applying parsimony to theory choice is one that (i) can include (NO Parsimony), (ii) definitely includes (KO-Parsimony) and (KI-Parsimony), but (iii) does not include (NI-Parsimony). Not coincidentally, this is a package that has recently gained prominence amongst metaphysicians who care about the parsimony of their theories. Even the

[8] Cf. Cowling (2013); Sider (2011).

[9] Some philosophers contest the connection between parsimony and truth—either as it relates to metaphysical theories specifically or as it relates to any descriptive theory. Cf. Brenner (2017); Sober (2015); Willard (2014).

doi: 10.48106/dial.v74.i4.04

most ardent supporters of parsimony have shied away from including (NI-Parsimony). Sider (2011, 258–259) admits that '[t]here *is* a real question about which of propositional logic's connectives carve at the joints, and similarly for ∀ and ∃,' and yet nevertheless 'egalitarian answers can be given. . . [o]ne might hold that both ∃ and ∀ carve at the joints, or that all the truth-functional connectives do, and thus avoid drawing invidious metaphysical distinctions.'[10]

But, as I will now show, there is a puzzle that undermines this package's credibility.

2 The Puzzle

In this section, I will argue that (KO-Parsimony) entails a restricted form of (NI-Parsimony). I will then suggest that this is no mere entailment; properly understood, (KO-Parsimony) and (NI-Parsimony) target the same feature of reality, the structure of the world. Thus, insofar as these two approaches to parsimony are motivated by a desire to posit a simple world, it is puzzling that metaphysicians should treat them so differently.

To illustrate these connections, I will work through a paradigm example of the neo-Quinean methodology at work in the metaphysics of composite objects.

According to compositional nihilism, there are no composite objects—no tables, no chairs, and no people (if people are composite objects). Yet natural language claims like

> Some composite objects are larger than other composite objects

seem undeniably true.[11] The most straightforward regimentation of this English claim using first order logic is:

$$\exists x \exists y (C(x) \wedge C(y) \wedge (x \neq y) \wedge L(x, y))$$

[10] Here, I avoid using truth-functional operators (like propositional logic's connectives) as examples, since someone may argue that truth-functional operators aren't primitive anyway. Instead, truth-functional operators may be defined in terms of their truth-tables, which ultimately depend on primitive notions of truth and falsity. Thanks to an anonymous reviewer for this suggestion.

[11] Some metaphysicians (e.g., Merricks 2001) say that such claims are false. Nevertheless, there is a sense in which such claims are "nearly as good as true." Nothing in what follows depends on the difference between what is true (and later paraphrased) and what is nearly as good as true. Cf. Bennett (2009, 58–59).

which informally reads 'There is an *x* and there is a *y* such that *x* is a composite object, *y* is a composite object, *x* is not identical with *y*, and *x* is larger than *y*.' According to orthodox neo-Quineanism, if we endorse this regimentation we thereby incur an ontological commitment to composite objects.

But we want to avoid an ontological commitment to composite objects. This is in part because (KO-Parsimony) recommends reducing the number of posited ontological kinds when feasible. Composite objects form an ontological kind. So we ought to avoid positing them.

How do we accomplish that goal? It is not enough merely to reduce the number of references to composite objects or to relegate claims about composite objects to a theoretically insignificant role. On the neo-Quinean methodology, we posit an ontological kind when, in stating our theory, we employ a predicate that ranges over entities found within that kind. Thus, we need to avoid the mention of composite objects altogether. To accomplish that, we need to find an alternative regimentation to the English sentence ('Some composite objects are larger than other composite objects.') that uses only nihilistically acceptable ideology.

Here's how we can do that. First, we replace the composite object predicate, '*C*,' with the predicate '*AC*,' which reads as 'arranged composite-object-wise.' This predicate ranges over the things that are spatially distributed as if they composed an object. If contemporary physics is correct, the entities that satisfy this predicate are quarks, leptons, and bosons. But so as to not presuppose any particular theory, let's call them—whatever they are—"simples." '*AC*' ranges over simples, but in a non-distributive manner. No single simple is arranged composite-object wise. Rather, all of the simples are collectively arranged composite-object-wise. Finally, we must be able to quantify over simples arranged composite-object-wise in a way that avoids committing ourselves to something "over and above" those simples. To that end, we supplement first-order logic's singular quantification with plural quantification. Following some fairly standard notation from Burgess and Rosen (1997), we can use doubled letters (e.g., '*xx*,' '*yy*') to represent the variables for plural quantification. We can then regiment the English sentence as follows:

$$\exists xx \exists yy (AC(xx) \wedge AC(yy) \wedge (xx \neq yy) \wedge L(xx, yy))$$

This sentence successfully avoids an ontological commitment to composite objects.

Yet things are not so simple. We can use plural quantification to eliminate singular references to composite objects. But English also plausibly includes

doi: 10.48106/dial.v74.i4.04

plural references to composite objects.[12] Consider, for example, the following sentence:

> Some composite objects are in contact only with one another.

We would need to employ plural quantification in the regimentation of this sentence even with an ontological commitment to composite objects. For instance, where '*T*' is a predicate that ranges over things in contact and '≺' is a special relation between individuals and pluralities of individuals, functioning like the English expression 'among':

$$\begin{gathered}\exists xx[\forall u((u \prec xx) \rightarrow C(u)) \land \\ \forall v \forall w(((v \prec xx) \land T(v,w)) \rightarrow ((w \prec xx) \land v \neq w))]\end{gathered}$$

From an ideological perspective, this regimented sentence is already quite ugly. But, because it employs a predicate for composite objects, it would commit us to the existence of composite objects. So, to avoid such a commitment, we must construct a different regimentation that does not use such a predicate. This nihilistically acceptable regimentation will be even uglier. That's because it must rely on plurally plural—i.e., perplural—quantification. Just as plural quantification ranges over pluralities of individuals, perplural quantification ranges over second-level pluralities of pluralities. Let's use tripled letters (e.g., '*xxx*,' '*yyy*') to represent the variables for perplural quantification. We then get the following regimentation:

$$\begin{gathered}\exists xxx[\forall uu((uu \prec xxx) \rightarrow AC(uu)) \land \\ \forall vv \forall ww(((vv \prec xxx) \land T(vv,ww)) \rightarrow ((ww \prec xxx) \land vv \neq ww))]\end{gathered}$$

In this way, metaphysicians can avoid an ontological commitment to composite objects, thereby minimizing the kinds of objects to which they are ontologically committed. But their use of (primitive) perplural quantification increases the ideological kinds to which they are committed.

So far as the metaphysics of composite objects goes, we have two options. First, we can employ a predicate that ranges over composite objects. Or, to avoid the ontological commitment, we can remove the predicate. Choosing this second option seems to involve a trade-off between a specially problematic predicate and a more complicated form of quantification.

[12] It is contentious whether English contains genuine perplural locutions (see Linnebo and Nicolas 2008; McKay 2006, 46–52). I cannot speak to other natural languages.

Our intuitive grasp of the relevant concepts initially suggested that ontology and ideology are quite distinct. So it's surprising that a commitment to (KO-Parsimony) entails a *de facto* commitment to (NI-Parsimony). This connection cries out for explanation.

In fact, I think the explanation is quite straightforward for ideological externalists. If we use a theory's ideology to pick out features of the world, then it's entirely plausible that in doing so we sometimes pick out ontological kinds.

Think of it this way. The elimination of a single object from a metaphysician's ontology improves its quantitative ontological parsimony. So, too, does the elimination of every object of a given kind. But the elimination of an ontological kind does not necessarily result in the elimination of any objects. It's perfectly ordinary for a reductive project to "relocate" the objects of one kind into the province of another. For example, David Lewis's modal realism (1986) is ontologically parsimonious insofar as it avoids an ontological commitment to *sui generis* possible worlds. But it does not minimize the overall number of objects; in a manner of speaking, what would have been *sui generis* possible worlds are instead causally isolated concrete entities. So, (KO-Parsimony) should not be understood as an efficient means of reducing the overall number of objects posited. Similarly, (KO-Parsimony) should not be understood merely as a preference for "empty kinds" over "populated kinds." In many cases, whether or not an ontological kind is populated should depend on contingent facts of the world rather than metaphysical necessities. (KO-Parsimony) should be understood as a preference for the *elimination* of ontological kinds. As the compositional example above suggests, the elimination of an ontological kind is achieved by the abandonment of its corresponding predicate. Here is where ideological externalism is relevant. When a theory commits to an ontological kind, it is not committing to some object that it quantifies over. Rather, when a theory commits to an ontological kind, it is committing to a structural feature of the world that corresponds to a predicate employed by the theory's ideology. Similarly, when a theory eliminates an ontological kind, it eliminates a structural feature of the world. Ontological kinds are features of the world's metaphysical structure.

Compare this theoretical identification to the theoretical identification of water and H_2O. Our concept of water is quite different from our concept of H_2O: our concept of water predates our concept of H_2O; our concept of water is rooted in its geographic, biological, and sociological functions whereas our concept of H_2O is rooted in the scientific discipline of chemistry; and so on.

As a matter of fact, though, the two concepts pick out the same substance. Of course, in some sense our concept of water "could have" picked out a different substance. Perhaps, even, our concept of water "could have" picked out a metaphysically gruesome disjunction of substances. But that's not how things turned out. Consequently, to be concerned with water is to be concerned with H_2O. Imagine someone who stressed the importance of bringing water on a camping trip. If they stressed the importance of bringing *water* but denied the importance of bringing H_2O, we would be confused—and rightly so.

So, too, for ontological kinds and the world's metaphysical structure. While our concept of an ontological kind may predate our concept of the world's metaphysical structure, the two concepts ultimately pick out the same feature. Of course, there may be some differences between the two theoretical identifications. Those who maintain a firm distinction between the *a priori* and the *a posteriori* would likely consider "Water is H_2O" to be an *a posteriori* identification and "Ontological kinds are metaphysical structure" to be an *a priori* identification. But, assuming the identities hold, many of the comparisons are apt. If a metaphysician stresses the importance of minimizing the ontological kinds posited by a theory, we should expect them to stress the importance of minimizing the structural complexity posited by a theory—it's the same thing that is being minimized! At a minimum, the metaphysician owes us an explanation for the difference in attitude.

Thus far, I have argued that those committed to (KO-Parsimony) should be committed to a restricted version of (NI-Parsimony). I have also suggested that there is an identity between the targets of these two principles of parsimony; both seek to minimize the structural complexity of the world. It does not follow that qualitative ontological parsimony *just is* quantitative ideological parsimony. There will still be instances of the latter that aren't instances of the former. Consider, for instance, a choice between two competing modal theories. Some forms of actualism (like those in Prior and Fine 1977) eschew quantifying over possible worlds and take the sentential modal operators as primitive. Suppose that actualist theory T_1 takes both '$\Box$' and '$\Diamond$' as primitive and actualist theory T_2 takes only '$\Box$' as primitive, defining '$\Diamond$' in the standard way. (NI-Parsimony) would recommend T_1 over T_2 because it employs one less bit of ideology. But by hypothesis neither theory posits more or fewer kinds of entities. Thus, some disputes about ideology are not reducible to disputes that involve ontology.[13]

[13] I develop this point more fully in Finocchiaro (2019b).

Here's a small, but important, complication that I've ignored.[14] Thus far, I have worked through a single case, the metaphysics of composition. Even if what I have said holds for this case, does the point generalize? Or is it merely an artifact of the case that might or might not apply to others?

The point generalizes. On the neo-Quinean paradigm, there is no ontological commitment to something unless there is a regimented sentence held to be true which includes a bound variable that must refer to that thing. But there is no need to have such a referring bound variable unless that variable attaches to a predicate of some kind. In other words, because ontological parsimony is a difference in ontology and because ontology is always expressed through ideology, ontological parsimony always involves a difference in ideology.

There is one slight exception. Some metaphysicians adopt principles of parsimony that discriminate on the basis of fundamentality. For example, Schaffer (2009) adopts the Laser, which recommends minimizing the number of *fundamental* entities but does not care about the number of *non-fundamental* entities. Such a principle makes the connection between ontology and ideology weaker. More specifically, when using the Laser there will be predicates—the ones corresponding to non-fundamental entities—whose elimination or introduction would not impact ontological parsimony.

But this exception does not solve the puzzle. First, it's unclear what the status of such predicates is. Plausibly, non-fundamental ontology is expressed through non-fundamental ideology. If so, then this exception is simply irrelevant to the puzzle I've presented. Second, this exception still entails a strong relationship between fundamental ontology and fundamental ideology. So, at best, it would solve only part of the puzzle.

3 What to Do?

I will end by briefly discussing three ways to respond to the puzzle about parsimony. Each has its advantages and disadvantages. While I do favor one of the ways over the others, I think all three are worth developing more fully.

First, we could try to resist the puzzle. I generated the puzzle by assuming orthodox neo-Quineanism. One way of resisting, then, is to reject the claim that a theory's ontology is that over which the theory quantifies. There are several alternatives to the Quinean criterion of ontological commitment, but one promising option is the truthmaker view. On the truthmaker view,

[14] Thanks to an anonymous reviewer for pushing me to address this issue.

a theory's ontology is that which makes the theory's sentences true.[15] Importantly, the view explicitly permits two theories to differ with respect to their ideologies without also differing with respect to their ontological commitments. For instance, on the truthmaker view a theory might truly state "Some composite objects are larger than other composite objects" without incurring an ontological commitment to composite objects. What matters is not what the sentence quantifies over but rather what makes the sentence true—and what makes the sentence true need not be composite objects. More importantly, the view entails that the two regimentations offered above—"$\exists x \exists y(C(x) \wedge C(y) \wedge (x \neq y) \wedge L(x,y))$" and "$\exists xx \exists yy(AC(xx) \wedge AC(yy) \wedge (xx \neq yy) \wedge L(xx,yy))$"—have the same ontological commitments. The change in ideology does not impact the ontology. Thus, on the truthmaker view of ontological commitment, (KO-Parsimony) does not entail any version of (NI-Parsimony), nor does it suggest an identity between their targets. In a way, then, the puzzle about parsimony could motivate us to reject orthodox neo-Quineanism.

Those of us not yet ready to abandon orthodoxy have to either embrace the puzzle or downplay its significance. I suspect many would prefer the second option. Some metaphysicians (e.g., Bennett 2009) have characterized many metaphysical disputes as being, as bottom, trade-offs between ontology and ideology. This characterization is hard to maintain if they have the same target (i.e., the world's metaphysical structure). It seems, then, that my puzzle puts that characterization in a hard place. But perhaps the essence of their characterization can be maintained. I can see two strategies for doing so.

On the first strategy, there are many more ideological kinds than previously assumed. More specifically, each predicate that expresses an ontological kind forms its own ideological kind. If this is so, then (KO-Parsimony) actually entails (KI-Parsimony) and the methodological tension vanishes. But here's a challenge that this strategy must overcome. By following the neo-Quinean orthodoxy, we eliminate ontologically committing predicates but we do not eliminate the complements of those predicates. So, for instance, the compositional nihilist eliminates 'composite object' but does not eliminate 'not a composite object,' otherwise known as 'simple.' Yet, intuitively, "positive"

[15] See Rettler (2016, 21). Rettler even appears to gesture toward a version of my puzzle when he says, "[I]t's true, just looking at the sentences will no longer tell you which theory wins the day with respect to parsimony of ontological commitments. But it never should have." In what follows I will simplify my discussion by ignoring Rettler's distinction between the general truthmaker view and the specific truthmaker view.

predicates like ‘composite object’ and “negative” complements like ‘simple’ are of the same ideological kind. So, those who want to pursue this first strategy of downplaying the significance of the puzzle must offer a more sophisticated means of individuating ideological kinds.

On the second strategy, there are two categories of ideology such that (i) we ought to minimize the number of ideological bits from the first category, and (ii) it is not the case that we ought to minimize the number of ideological bits from the second category. Obviously, those who pursue this strategy must offer some explanation for the difference in treatment. One somewhat radical explanation is to say that structural simplicity is more important in some domains than it is in others. I do not see how this explanation can be plausibly maintained. Parsimony is currently treated as a comprehensive value: choose the theory that is *overall* more simple. Why would simplicity in one domain be less important (i.e., less truth conducive) than simplicity in another domain? On an alternative explanation, the relationship between ideological bits and metaphysical structure is more nuanced than previously thought. Perhaps ideological bits are more fine-grained than the corresponding structure. If so, then some ideological bits (like ‘is a part of’ and ‘overlaps’) would correspond to the same aspect of the world’s metaphysical structure, and so there is no need to choose between the two. In contrast, other ideological bits (like ‘composite object’ and ‘simple’) would correspond to different aspects of the world’s metaphysical structure, and so there is value in eliminating one if not the other. This explanation is interesting. But as it stands it is *ad hoc*. In the absence of a worked-out account of ideological correspondence, why should we think that it works the way this strategy needs it to work?

That leaves the third response: embrace the puzzle. If we embrace the puzzle, we ought to claim that (NI-Parsimony) is no less justified a principle than (KO-Parsimony). This claim is quite shocking (well, as shocking as an esoteric claim about the proper methodological application of parsimony can be, anyway). (KO-Parsimony) has a rich history and is likely the most broadly endorsed approach to parsimony. In contrast, almost no one explicitly endorses (NI-Parsimony). Nevertheless, by embracing the puzzle we can save neo-Quineanism as well as the standard characterization of metaphysical disputes as disputes that involve trade-offs between ontology and ideology. Yet those who pursue this third strategy have their own explaining to do. Intuitively, it seems objectionably arbitrary to choose between functionally equivalent terminology. So why isn’t it? For example, why should we reduce the number of compositional predicates we employ in stating our theories?

Perhaps we can extend the standard motivations for parsimony-based considerations and say that we should reduce the number of compositional predicates because the resulting theory posits a more simple structure and is therefore more likely to be true. This might still generate an epistemic deadlock with regard to competing “equivalent” theories. (NI-Parsimony) would suggest that a theory that employs only ‘overlaps’ is more likely to accurately represent the compositional structure of the world than a theory that employs both ‘overlaps’ and ‘is a part of.’ *Mutatis mutandis* for a theory that employs only ‘is a part of.’ But at this point our methodology fails us and we do not know which of the two predicates we ought to employ.[16]

Personally, I think we ought to embrace the puzzle. It’s not a perfect response, but it is the best available. Neo-Quineanism is battle-tested orthodoxy. (More modestly, neo-Quineanism is much closer to the center of my web of belief than are the other elements of the puzzle.) For that reason I reject the first response. The second response raises a number of issues regarding ideological correspondence. I am doubtful that those issues can be addressed satisfactorily. So I also reject the second response. Finally, I do not think that the third response is that bad. I don’t know how to choose between overlap and parthood. I don’t even know how to think about that choice. But a hard choice is not *ipso facto* a bad choice.[*]

Peter Finocchiaro
0000-0003-4060-7061
Wuhan University
peter.w.finocchiaro@gmail.com

References

Bennett, Karen. 2009. “Composition, Colocation, and Metaontology.” in *Metametaphysics. New Essays on the Foundations of Ontology*, edited by David J. Chalmers, David Manley, and Ryan Wasserman, pp. 38–76. Oxford: Oxford University Press, doi:10.1093/oso/9780199546046.001.0001.

Brenner, Andrew. 2017. “Simplicity as a Criterion of Theory Choice in Metaphysics.” *Philosophical Studies* 174(11): 2687–2707, doi:10.1007/s11098-016-0805-1.

[16] Cf. McSweeney (2019, 127–128).

[*] For their help in the development of this paper, I thank Andrew Brenner, Rebecca Chan, Alessandro Torza, Jeffrey J. Watson, Qiong Wu, the audience at my 2020 Central APA conference session, and two anonymous reviewers. I also thank Collegetown Bagels for their bagels, coffee, and outdoor seating, all of which are indispensable for writing during a pandemic.

BURGESS, John P. and ROSEN, Gideon. 1997. *A Subject With No Object: Strategies for Nominalistic Interpretations of Mathematics*. Oxford: Oxford University Press, doi:10.1093/0198250126.001.0001.

CAMERON, Ross P. 2012. "Why Lewis's Analysis of Modality Succeeds in Its Reductive Ambitions." *Philosophers' Imprint* 12(8).

CHALMERS, David J., MANLEY, David and WASSERMAN, Ryan, eds. 2009. *Metametaphysics. New Essays on the Foundations of Ontology*. Oxford: Oxford University Press, doi:10.1093/oso/9780199546046.001.0001.

CHURCHLAND, Paul M. 1984. *Matter and Consciousness*. Cambridge, Massachusetts: The MIT Press.

COWLING, Sam. 2013. "Ideological Parsimony." *Synthese* 190(17): 3889–3908, doi:10.1007/s11229-012-0231-7.

FINOCCHIARO, Peter. 2019a. "The Explosion of Being: Ideological Kinds in Theory Choice." *The Philosophical Quarterly* 69(276): 486–510, doi:10.1093/pq/pqz005.

—. 2019b. "The Intelligibility of Metaphysical Structure." *Philosophical Studies* 176(3): 581–606, doi:10.1007/s11098-017-1029-8.

—. 2021. "Ideology and Its Role in Metaphysics." *Synthese* 198(2): 957–983, doi:10.1007/s11229-018-02077-6.

HORGAN, Terence E. and POTRČ, Matjaž. 2008. *Austere Realism: Contextual Semantics Meets Minimal Ontology*. Cambridge, Massachusetts: The MIT Press, doi:10.7551/mitpress/9780262083768.001.0001.

LEWIS, David. 1973. *Counterfactuals*. Cambridge, Massachusetts: Harvard University Press. Cited after republication as Lewis (2001).

—. 1983. *Philosophical Papers, Volume 1*. Oxford: Oxford University Press, doi:10.1093/0195032047.001.0001.

—. 1986. *On the Plurality of Worlds*. Oxford: Basil Blackwell Publishers.

—. 2001. *Counterfactuals*. Oxford: Basil Blackwell Publishers. Republication of Lewis (1973).

LEWIS, David and LEWIS, Stephanie R. 1970. "Holes." *Australasian Journal of Philosophy* 48(2): 206–212. Reprinted in Lewis (1983, 3–9), doi:10.1080/00048407012341181.

LINNEBO, Øystein and NICOLAS, David. 2008. "Superplurals in English." *Analysis* 68(3): 186–197, doi:10.1093/analys/68.3.186.

MCKAY, Thomas J. 2006. *Plural Predication*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199278145.001.0001.

MCSWEENEY, Michaela Markham. 2019. "Following Logical Realism Where It Leads." *Philosophical Studies* 176(1): 117–139, doi:10.1007/s11098-017-1008-0.

MELIA, Joseph. 2008. "A World of Concrete Particulars." in *Oxford Studies in Metaphysics*, volume IV, edited by Dean W. ZIMMERMAN, pp. 99–124. Oxford: Oxford University Press.

Merricks, Trenton. 2001. *Objects and Persons*. Oxford: Oxford University Press, doi:10.1093/0199245363.001.0001.

Nolan, Daniel Patrick. 1997. “Quantitative Parsimony.” *The British Journal for the Philosophy of Science* 48(3): 329–343, doi:10.1093/bjps/48.3.329.

Parsons, Josh. 2014. “The Many Primitives of Mereology.” in *Mereology and Location*, edited by Shieva Kleinschmidt, pp. 3–12. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199593828.001.0001.

Paul, Laurie A. 2017. “A One Category Ontology.” in *Being, Freedom, and Method: Themes from the Philosophy of Peter van Inwagen*, edited by John Adorno Keller, pp. 32–61. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780198715702.001.0001.

Prior, Arthur Norman and Fine, Kit. 1977. *Worlds, Times and Selves*. London: Gerald Duckworth & Co. Edited by Kit Fine; based on manuscripts by Prior with a preface and a postscript by Kit Fine.

Quine, Willard van Orman. 1948. “On What There Is.” *The Review of Metaphysics* 2(5): 21–38. Republished as Quine (1951), reprinted in Quine (1953, 1–19).

—. 1951. “On What There Is.” *Proceedings of the Aristotelian Society, Supplementary Volume* 25: 217–234. Reprint of Quine (1948), doi:10.1093/aristoteliansupp/25.1.125.

—. 1953. *From a Logical Point of View: 9 Logico-Philosophical Essays*. Cambridge, Massachusetts: Harvard University Press. Cited after the revised edition: Quine (1961).

—. 1961. *From a Logical Point of View: 9 Logico-Philosophical Essays*. 2nd ed. Cambridge, Massachusetts: Harvard University Press. Revised edition of Quine (1953), reprinted 1980.

Rettler, Bradley. 2016. “The General Truthmaker View of Ontological Commitment.” *Philosophical Studies* 173(5): 1405–1425, 1427, doi:10.1007/s11098-015-0526-x.

Schaffer, Jonathan. 2009. “On What Grounds What.” in *Metametaphysics. New Essays on the Foundations of Ontology*, edited by David J. Chalmers, David Manley, and Ryan Wasserman, pp. 347–383. Oxford: Oxford University Press, doi:10.1093/oso/9780199546046.001.0001.

—. 2015. “What Not to Multiply Without Necessity.” *Australasian Journal of Philosophy* 93(4): 644–664, doi:10.1080/00048402.2014.992447.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199697908.001.0001.

—. 2013. “Against Parthood.” in *Oxford Studies in Metaphysics*, volume VIII, edited by Karen Bennett and Dean W. Zimmerman, pp. 237–293. New York: Oxford University Press, doi:10.1093/acprof:oso/9780199682904.001.0001.

Sober, Elliott R. 2015. *Ockham’s Razors. A User’s Manual*. Cambridge: Cambridge University Press, doi:10.1017/CBO9781107705937.

TALLANT, Jonathan. 2013. “Quantitative Parsimony and the Metaphysics of Time: Motivating Presentism.” *Philosophy and Phenomenological Research* 87(3): 688–795, doi:10.1111/j.1933-1592.2012.00617.x.

TURNER, Jason. 2015. “What’s So Bad About Second-Order Logic?” in *Quantifiers, Quantifiers, and Quantifiers: Themes in Logic, Metaphysics and Language*, edited by Alessandro TORZA, pp. 463–488. Synthese Library n. 373. Dordrecht: Springer Verlag, doi:10.1007/978-3-319-18362-6.

VAN INWAGEN, Peter. 1998. “Meta-Ontology.” *Erkenntnis* 48(2–3): 233–250. Reprinted in van Inwagen (2001, 13–31), doi:10.1023/A:1005323618026.

—. 2001. *Ontology, Identity, and Modality. Essays in Metaphysics*. Cambridge: Cambridge University Press.

WILLARD, Mary Beth. 2014. “Against Simplicity.” *Philosophical Studies* 167(1): 165–181, doi:10.1007/s11098-013-0228-1.

doi: 10.48106/dial.v74.i4.04

David Armstrong on the Metaphysics of Mathematics

Thomas M.E. Donaldson

This paper has two components. The first, longer component (sections 1–6) is a critical exposition of Armstrong's views about the metaphysics of mathematics, as they are presented in *Truth and Truthmakers* and *Sketch for a Systematic Metaphysics*. In particular, I discuss Armstrong's views about the nature of the cardinal numbers, and his account of how modal truths are made true. In the second component of the paper (section 7), which is shorter and more tentative, I sketch an alternative account of the metaphysics of mathematics. I suggest we insist that mathematical truths have physical truthmakers, without insisting that mathematical objects themselves are part of the physical world.

A prime number p is a "Sophie Germain prime" if $(2p + 1)$ is also prime. It is conjectured that there exist infinitely many Sophie Germain primes. I don't know whether this conjecture is true, but what I do know is that there exist *some* Sophie Germain primes: 2 is an example; 3 is another; $(2,618,163,402,417 \times 2^{1290000} - 1)$ is a third, or so I am told. Now it is obvious that every Sophie Germain prime is a number; and it follows that there exist some numbers—or so it seems.

But what kind of a thing is a number?

This is a difficult question, but one point at least seems clear: numbers and other mathematical entities are "abstract," in the sense that they have no causal powers and no location in spacetime. We are told that the number zero was discovered in India, but it would be a mistake to go to India *now* to look for it—and not because it has subsequently been moved. You can't trip over the number three. The polynomial $(x^2 - 3x + 2)$ can be split into two factors, $(x - 2)$ and $(x - 1)$, but not by firing integers at it in a particle accelerator. The empty set has no gravitational field. And so on.

And so, to labour the point, it seems that some abstract entities exist.

And yet David Armstrong began his last book by endorsing what he called "naturalism":

> I begin with the assumption that all that exists is the space-time world, the physical world as we say. [...] [This] means the rejection of what many contemporary philosophers call "abstract objects", meaning such things as numbers or Platonic Forms or classes, where these are supposed to exist "outside of" or "extra to" space-time. (2010, 1)

Despite this naturalism, Armstrong did not reject any part of mainstream mathematics. Indeed, he insisted that the truths of orthodox pure mathematics are necessary and *a priori* (2010, ch. 12).

In this paper, I explore Armstrong's attempt to reconcile his denial of abstract entities with his commitment to orthodox mathematics. To be more specific, the paper has three goals.

1. Armstrong wrote a vast amount on mathematics, and this writing is spread among many papers and books. Some of this work is complex, and Armstrong changed his mind on certain important questions. My first goal in this paper is to describe—clearly, briefly, and in one place—Armstrong's mature views on the metaphysics of mathematics, including relevant aspects of his work on the metaphysics of modality. To prevent the discussion from sprawling, I focus particularly on Armstrong's account of cardinal number, as it is presented in his last two books: *Truth and Truthmakers* (2004, "T&T") and *Sketch for a Systematic Metaphysics* (2010, "SSM"). Section 1 and section 2 describe Armstrong's views about the cardinal numbers; sections 3–6 focus on modality.
2. My second goal is to present some novel—and, I believe, definitive—objections to Armstrong's views on the metaphysics of mathematics. These objections are presented in sections 5 and 6.
3. My third goal is to recommend a different way of thinking about the metaphysics of mathematics—an approach which will, I hope, appeal to people who admire Armstrong's work. Briefly, I will suggest that we insist that every mathematical truth has a truthmaker in the physical world, without also insisting that mathematical objects themselves are physical things. The proposal is presented in more detail in section 7.

1 Cardinal Numbers as Concrete Entities

The claim that numbers have no spatial location is familiar to metaphysicians, but it sometimes comes as a surprise to students. “But there are three pens on my desk *right now!*” they exclaim, implying that the number three itself is within arm’s reach. According to Armstrong, the surprised students are on to something.

While he rejected “Platonic forms” which exist outside spacetime, Armstrong did believe that there are properties which exist *within* the particulars that instantiate them (SSM, ch. 2). For Armstrong, there exists a property *is red* which exists within London buses, ripe tomatoes, and male cardinals. As he sometimes put it, properties are “immanent” rather than “transcendent.” He inferred that all properties are instantiated. Uninstantiated properties have no place in spacetime, and so no place in Armstrong’s philosophical system (SSM, 15–16). He made the same claim about relations (SSM, 23).

For Armstrong, a cardinal number is a relation between a particular and a property. Specifically, the cardinal number $\varkappa$ is a relation that a particular x bears to a property P just in case x has, as mereological parts, exactly $\varkappa$ particulars which instantiate P. A normal octopus bears the *one* relation to the property *is an octopus,* and the *eight* relation to the property *is a limb.* The mereological sum of two normal octopuses bears the *two* relation to the property *is an octopus* and the *sixteen* relation to the property *is a limb.* And so on.[1, 2]

On Armstrong’s view, then, the surprised student is correct to think that cardinal numbers exist within the “spacetime world.” It turns out, then, that one can coherently maintain that numbers exist and that there are no abstract entities.

This is a striking result—and yet, a problem looms. As I said, Armstrong insisted that properties and relations exist only when they are instantiated. On this view, the number $10^{10^{10}}$ exists only if the spacetime world happens to

[1] Armstrong’s theory of cardinal number is closely related to that presented in Kessler (1980). Simons (1982) raises some important objections to Kessler’s account. I believe that Armstrong’s theory is not vulnerable to Kessler’s objections, though I will not pursue the issue here.

[2] Note that the property P will in many cases be a “second-rate property” rather than a genuine universal (SSM, 19). For example, there are seven medium-sized red spoons in my apartment. As Armstrong would put it, the fusion of the contents of my apartment bears the *seven* relation to the property *medium-sized red spoon.* But Armstrong would surely deny that *medium-sized red spoon* is a universal. Armstrong understood that we need “second-rate” properties in cases like this (T&T, 72).

doi: 10.48106/dial.v74.i4.05

contain $10^{10^{10}}$ particulars. And it is far from clear that the spacetime world is that large. I call this “the problem of size.”

It is tempting to reply to this objection by insisting that space is infinitely divisible. Discussing the problem of size as it arises for Aristotle, Jonathan Barnes writes:

> Physical objects are, in Aristotle’s view, infinitely divisible. That fact ensures that, even within the actual finite universe, we shall always be able to find a group of k objects, for any k [...] If the universe consisted simply of a single sphere, it would also contain two objects (two hemispheres), three objects (three third-spheres) and so on. We shall never run short of numbers of things [...]. (1985, 122)

This cannot be considered a satisfactory solution to *Armstrong’s* problem, however. For one thing, it is far from clear that Aristotle was correct in thinking that space is infinitely divisible: those who study quantum gravity have been known to speculate that space is in fact discrete. And so the proposed solution is somewhat “hostage to fortune.” More importantly, even if the proposed account does secure the existence of large finite numbers such as $10^{10^{10}}$, it still leaves the existence of transfinite cardinals open to doubt. Standard mathematical descriptions of spacetime (which do imply infinite divisibility) entail that the set of spacetime points has cardinality $\beth_1$. Such accounts leave it unclear whether larger cardinal numbers (e.g., $\beth_2$, $\beth_3$, or even $\beth_\omega$) are instantiated in the physical world—and yet these larger cardinals are very much a part of orthodox mathematics.[3]

In passing, I note that Armstrong faced a problem of size in his account of set theory too. While Armstrong identified cardinal numbers with relations, he identified sets with individuals. The central claim in Armstrong’s account was David Lewis’s “brilliant insight” (T&T, 120): the mereological parts of a set are precisely its non-empty subsets. For example, Lewis’s claim implies that the mereological proper parts of $\{a, b, c\}$ are $\{a\}$, $\{b\}$, $\{c\}$, $\{a, b\}$, $\{b, c\}$, and $\{c, a\}$. This implies that every singleton set is mereologically simple. Now the set theorists tell us that for each cardinal number $\varkappa$, there are at least $\varkappa$ singletons. Thus, Armstrong is stuck with the claim that, within the spacetime world, there are least $\varkappa$ mereologically simple individuals, for each $\varkappa$. And this seems highly doubtful. Perhaps one could plausibly argue that there are $\beth_1$

[3] For an introduction to the mathematics of these enormous numbers, see Yarnelle (1964).

mereological simples by saying that each spacetime point is mereologically simple and there are $\beth_1$ of those. But it is hard to see any justification for the claim that there are $\beth_2$, or $\beth_3$, or even $\beth_\omega$ mereological simples in the spacetime world. Once again, the physical universe seems to be too small to accommodate the ontology of mathematics.[4]

2 Armstrong's Possibilism

Armstrong was aware of the problem of size. He responded by claiming that larger cardinal numbers exist *in posse* though not *in esse:*

> A Platonist will solve the problem [of size] by postulating uninstantiated numbers. [...] [However,] my own hard choice [...] is to accept a deflationary doctrine of what it is for a mathematical entity to exist. Plenty of mathematical structures exist in a straightforward sense, because they are instantiated. We can call them *empirical* mathematical structures. [...] But mathematical existence itself, I suggest, should be reckoned as something less. A mathematical entity exists if and only if it is *possible* that there should be instantiations of that structure. (T&T, 117)

Armstrong called this *possibilism.* The doctrine is not easy to interpret. In the above passage, Armstrong seems to suggest that a mathematical entity exists provided that it *could* be instantiated—even if it is in fact not instantiated. However, in this same passage, Armstrong *contrasts* his possibilism with the "Platonist" claim that there are uninstantiated mathematical entities.

The following quotation gives us a clue about what Armstrong meant:

> We say '$7 + 5 = 12$', but this can be rendered more transparently, though more boringly, as ⟨necessarily, if there are seven things and five further things, then the sum of these things are twelve things⟩. (T&T, 101)

What this passage suggests is that, for Armstrong, while the sentence "$7 + 5 = 12$" appears to describe a relation among three mathematical entities (viz. the numbers seven, five and twelve) it is in fact a generalization about pluralities of marbles, pebbles, sticks, or whatever. More generally, the possibilist maintains that while pure mathematics appears to describe a domain of special

[4] For a more thorough discussion of this point, see Rosen (1995).

doi: 10.48106/dial.v74.i4.05

mathematical entities, it in fact consists of *modal* statements—statements about what is necessary or possible.

Now Armstrong did not develop this proposal systematically; instead, he endorsed Geoffrey Hellman's modal structuralism:

> I recognize, of course, that asserting here this [...] doctrine of mathematical existence is to a degree a matter of hand-waving. I have not the logico-mathematical grasp to defend it in any depth. That has been done, in particular by Geoffrey Hellman. (T&T, 117)

The reader who wants a thorough discussion of modal structuralism should consult Hellman (1989). For now, a back-of-the-envelope summary will be sufficient. The modal structuralist claims that the theory of the natural numbers is not a description of some particular sequence of entities; rather, the theory concerns *all possible* models of the Peano axioms.[5]

For example, when a mathematician asserts that there are infinitely many prime numbers, what is really meant is something like this:[6]

> It is necessary that, in any model of the Peano axioms, the domain contains infinitely many prime elements.

Notice that this statement does not imply that there is a model of the Peano axioms somewhere in spacetime.

When a modal structuralist mathematician asserts that, necessarily, every model of the Peano axioms contains infinitely many prime elements, she will of course wish to rule out the suggestion that this is true "vacuously"—that is, simply because models of the Peano axioms are *impossible*. Thus, a modal structuralist mathematician will claim that models of the Peano axioms are possible.

Hellman discusses in some detail how to extend this approach beyond number theory, and into applied mathematics. We need not look into these details. For our purposes, the key point is that the appeal to modal structuralism allows Armstrong to say that "$10^{10^{10}}$ is even" is true (when properly interpreted)

[5] The Peano axioms are the standard axioms in the theory of the natural numbers. Among them are such claims as "Zero is a number," and "If x is any natural number, $x + 0 = x$."

[6] The version of modal structuralism that I so quickly sketch here is hermeneutic rather than revolutionary (for this distinction, see Burgess, J. P. and Rosen 1997, 6–7). This is, I think, the correct interpretation of Armstrong's position. For Hellman's position, see (1998).

without committing himself to the questionable thesis that $10^{10^{10}}$ physical objects exist.[7]

The attractions of the approach are obvious, but Armstrong's possibilism brings with it a new problem. Armstrong was a truthmaker maximalist—he believed that every true proposition has a "truthmaker," that is, an entity in the spacetime world which is sufficient (and perhaps more than sufficient) to explain the truth of the proposition.[8] Armstrong was thus stuck with the formidable task of identifying truthmakers for complex modal truths like those described above. It is my contention that Armstrong did not succeed at this task, as I shall explain in the next three sections.[9]

3 Armstrong's Entailment Principle

Before we look at Armstrong's discussion of truthmaking and modality, we must consider his Entailment Principle, which is crucial to his account. Some notation will be helpful: I will put a sentence between angled brackets to represent the corresponding proposition. For example, ⟨Sam is dancing⟩ is the proposition that Sam is dancing. Here is the Entailment Principle, as it is formulated in *Sketch for a Systematic Metaphysics*:

> Entailment Principle (SSM Version). If α entails β, then any truthmaker for $\langle\alpha\rangle$ must be a truthmaker for $\langle\beta\rangle$ too. (SSM, 65–66)

For example, since φ entails $\neg\neg\varphi$, it is a consequence of Armstrong's Entailment Principle that any truthmaker for $\langle\varphi\rangle$ must also be a truthmaker for

[7] Hellman's modal structuralism involves second-order quantification, and it is worth thinking about how such quantifiers should be interpreted within Armstrong's metaphysical system. One approach is to say that the second-order variables range over properties (including "second-rate" properties—see footnote 2). Some restriction of the usual comprehension axiom will be needed to accommodate Armstrong's contention that there are no uninstantiated properties. For a version of modal structuralism that does not require second order quantification, see Berry (2018).

[8] The parenthetical "and perhaps more than sufficient" is there to indicate that Armstrong's was an *inexact* conception of truthmaking, to use Kit Fine's terminology—see (2017).

[9] Fox (1987) endorses a *purely modal* conception of truthmaking. According to Fox, T is a truthmaker for p just in case it is necessary that if T exists then p is true. On this approach, it is *easy* to identify truthmakers for purely mathematical truths. Since the truths of pure mathematics are necessary, given Fox's purely modal conception of truthmaking, *anything whatever* is a truthmaker for any purely mathematical truth.

Armstrong himself vigorously rejected this approach, insisting that truthmakers must be *relevant* to the propositions they make true (T&T, 11). For more on this theme, see Cameron (2018).

doi: 10.48106/dial.v74.i4.05

$\langle\neg\neg\varphi\rangle$. In this example, the Entailment Principle is plausible. However, there is an important objection to this formulation of the principle. As Restall (1996) has pointed out, this simple version of the Entailment Principle conflicts with a popular and appealingly simple (though not undisputed) account of truthmaking and disjunction:

> Disjunction Principle. T makes true the proposition $\langle\varphi \vee \psi\rangle$ if and only if T makes true $\langle\varphi\rangle$, or T makes true $\langle\psi\rangle$, or both.

To see the conflict, consider the following argument:

> Let φ and ψ be any two true sentences, and suppose that T is a truthmaker for $\langle\varphi\rangle$. Then since φ entails $(\psi \vee \neg\psi)$, T must also be a truthmaker for $\langle\psi \vee \neg\psi\rangle$. By the Disjunction Principle, T must be a truthmaker either for $\langle\psi\rangle$ or for $\langle\neg\psi\rangle$. But by hypothesis, $\langle\psi\rangle$ is true so $\langle\neg\psi\rangle$ is false and so $\langle\neg\psi\rangle$ has no truthmakers. So T must be a truthmaker for $\langle\psi\rangle$.

This little argument appears to show for any two true sentences φ and ψ, any truthmaker for $\langle\varphi\rangle$ is also a truthmaker for $\langle\psi\rangle$—a result which completely trivializes truthmaker theory.

In *Truth and Truthmakers,* Armstrong gives a more sophisticated version of the Entailment Principle which is not subject to the same objection:

> Entailment Principle (T&T Version). If α entails* β, then any truthmaker for $\langle\alpha\rangle$ must be a truthmaker for $\langle\beta\rangle$ too. (T&T, 10)

Here, entailment* is some non-classical entailment relation, to be specified. By insisting that φ need not entail* $(\psi \vee \neg\psi)$, we can maintain a version of the Entailment Principle without having to conclude, absurdly, that all propositions expressed by true sentences have the same truthmakers.[10]

We have seen that the SSM Version of the Entailment Principle conflicts with the Disjunction Principle, and that one can maintain the Disjunction Principle by endorsing the T&T Version instead. If a truthmaker theorist wishes instead to maintain the simpler, SSM Version of the Entail-

[10] It is not easy to say exactly what entailment* is. Restall (1996) has proposed that entailment* "is nearly, but not quite, the first degree entailment of relevant logic." Linnebo (2022) has suggested that entailment* includes first-order intuitionistic entailment without identity. Thankfully, we need not settle this question here.

ment Principle, she may choose to reject the Disjunction Principle—and indeed philosophers have presented *independent* reasons for rejecting this principle.[11] We need not settle this dispute here. Suffice it to say that appeals to the SSM Version of the Entailment Principle are subject to dispute.[12]

4 Armstrong on Truthmaking and Possibility

Having briefly looked at the Entailment Principle we are ready to consider Armstrong's account of truthmaking and modality. Let's start with possibility. Suppose that the sentence φ expresses a contingently true proposition; what then are the truthmakers for $\langle\Diamond\varphi\rangle$ and $\langle\Diamond\neg\varphi\rangle$?

$\langle\Diamond\varphi\rangle$ is comparatively straightforward. Since $\langle\varphi\rangle$ is true, Armstrong argued, it must have a truthmaker, T. Since φ entails $\Diamond\varphi$, T will be a truthmaker for $\langle\Diamond\varphi\rangle$ as well, by the Entailment Principle.[13]

$\langle\Diamond\neg\varphi\rangle$ is rather more difficult. Armstrong introduced his "possibility principle" (T&T, 84) to deal with the problem:

> Possibility Principle. If $\langle\varphi\rangle$ is a contingent truth and T is a truthmaker for $\langle\varphi\rangle$, then T is a truthmaker for $\langle\Diamond\neg\varphi\rangle$.

The principle is not attractive on its face. As Pawl (2010) has pointed out, Armstrong's being legged is a truthmaker for ⟨Someone has legs⟩, but it is hardly plausible that Armstrong's being legged is a truthmaker for ⟨Possibly, nobody has legs⟩. But Armstrong claimed that the Possibility Principle is a consequence of the Entailment Principle. He presented the following argument (T&T, 84, notation slightly modified):

(1) T is a truthmaker for $\langle\varphi\rangle$. (Assumption)
(2) $\langle\varphi\rangle$ is contingent. (Assumption)

[11] For discussion of the Disjunction Principle, see Rodriguez-Pereyra (2006) and López de Sa (2009).

[12] Why did Armstrong give these two different versions of the Entailment Principle, in books written in the same decade? My hypothesis is that T&T contains Armstrong's preferred formulation of the Entailment Principle, and that the version in SSM (a much shorter, easier work) is a simplification.

[13] It is not entirely straightforward that this application of the Entailment Principle is correct. While it is clear that φ entails $\Diamond\varphi$, it is perhaps not so clear that φ entails* $\Diamond\varphi$. We can let this point slide, however, because there are much more serious objections to Armstrong's position, as we shall see.

(3) $\langle\varphi\rangle$ entails $\langle\Diamond\neg\varphi\rangle$. (From (2), and the nature of the contingency of propositions)

(4) T is a truthmaker for $\langle\Diamond\neg\varphi\rangle$. (From (1), (3) and the Entailment Principle)

As Armstrong later recognized, this argument is fallacious. The error is in step (3): in no standard modal logic is it true that $\langle\varphi\rangle$ entails $\langle\Diamond\neg\varphi\rangle$ (special cases aside).

So the Possibility Principle is implausible on its face, and the argument Armstrong gave for it is unconvincing. I think we should conclude that the principle should be rejected.[14]

Happily, Armstrong also offered another and more attractive account of truthmaking and possibility. The idea is this. Suppose that we have (separately) two slices of bread, fifteen slices of cheese, and two slices of tomato. These things could have constituted a cheese and tomato sandwich—although in fact they don't. Plausibly, they together form a truthmaker for the proposition that a cheese and tomato sandwich *could* exist. Armstrong wrote:

> Consider, in particular, the cases where the entities in question do not exist, where they are *mere* possibilities. It is, let us suppose, true that ⟨it is possible that a unicorn exists⟩. What then is a minimal truthmaker for this truth? The obvious solution is combinatorial. The non-existent entity is some non-existent (but possible) combination out of elements that do exist. The phrase "non-existent combination" may raise eyebrows. Am I committing myself to a Meinongian view? No, I say. The *elements* of the combination are, I assert, the only truthmakers that are needed for the truth that this combination is possible. (T&T, 91–92)

I think that this is a very attractive account of what the truthmakers are for *some* truths about possibility (including truths about unicorns and tomato sandwiches).[15] However, it is doubtful that the combinatorial approach provides us with sufficient truthmakers for all the possibility claims made by the modal structuralist mathematician. The modal structuralist will assert

[14] Armstrong (2007) recognized the error in his argument for the Possibility Principle and went on to offer a *new* argument for it. For criticism of this later argument, see Pawl (2010).

[15] For some criticisms of Armstrong's "combinatorialist" theory of modality, see Wang (2013).

that second-order ZFC could have had a model, but it seems unlikely that such a model could be created by recombining physical objects, because it seems unlikely that there are *enough* physical objects to go around. If, for example, there are only $\beth_3$ physical objects, we will not by combining them be able to produce a set with $\beth_4$ elements, but *every* model of second-order ZFC contains sets with $\beth_4$ elements—and indeed much larger sets to boot. The problem of size has reemerged in a new form.

5 Armstrong on Truthmaking and Necessity (Part 1: Truth and Truthmakers)

Let's turn to Armstrong's discussion of propositions about necessity. Since our concern is Armstrong's philosophy of mathematics, we need not discuss all of Armstrong's views about truthmaking and necessity. Instead, we'll focus on what he had to say about truthmakers for the theorems of mathematics. Armstrong's discussions of this topic in *Truth and Truthmakers* and *Sketch for a Systematic Metaphysics* are very different. In this section, I'll consider chapter eight of *Truth and Truthmakers,* leaving the later book until section 7.

Armstrong (T&T, 99, 111) suggested that the numbers themselves constitute truthmakers for some arithmetical truths. For example, seven, five and twelve may together form a truthmaker for $\langle 7 + 5 = 12 \rangle$.[16] For Armstrong, so long as seven, five and twelve exist they *must* be related in this way, and so nothing *beyond* their existence is needed to explain their being so related. This relation between the three numbers is "internal" to them.

This is an important idea, and I will return to it in section 6. But this is not on its own a complete solution to the problem at hand. Consider for example the proposition $\langle \beth_\omega + \beth_\omega = \beth_\omega \rangle$. This is a theorem of orthodox mathematics, and so Armstrong would surely accept that the proposition is true, when given its proper modal interpretation. But what is its truthmaker? Surely $\beth_\omega$ itself can be a truthmaker only if it exists. However, as we saw in section 1 and section 2, it is doubtful for Armstrong that $\beth_\omega$ exists.[17]

Later in the chapter, Armstrong discussed analytic truths. He wrote:

[16] A variant on this suggestion (T&T, 98) is that any truthmaker for ⟨7 exists⟩, ⟨5 exists⟩ and ⟨12 exists⟩ will also be a truthmaker for $\langle 7 + 5 = 12 \rangle$.

[17] Of course, for Armstrong, "$\beth_\omega$ exists" is true when given a suitable modal reinterpretation—but this is only because, so interpreted, the sentence doesn't actually assert the existence of $\beth_\omega$!

> A traditional view, which has many supporters, is that [analytic] truths are true solely in virtue of the meanings of the terms in which they are expressed. (T&T, 109)

Armstrong went on to say that "[t]he phrase 'in virtue of' inevitably suggests truthmakers." So Armstrong proposed that if a sentence S is analytic, the proposition it expresses is made true by the meanings of the words in S. For example, ⟨A father is a male parent⟩ is made true by the meanings of "a," "father," "is," "a," "male" and "parent."

Now Armstrong suggested—somewhat tentatively—that statements in mathematics about what is necessary are analytic.[18] On this view, the meanings of mathematical terms make true all such statements.

I do not dismiss completely the claim that mathematical truths are analytic.[19] However, Armstrong's version of this thesis is insufficient to solve the problem at hand. Let *p* be some true proposition from pure mathematics. Armstrong believed that the theorems of pure mathematics are *necessarily* true. So we can ask what the truthmaker for *p would* have been, had there been no language-users. How might Armstrong reply? Surely it is not adequate to say that the meanings of English words would have been the truthmakers—for English words would not have existed in the absence of English speakers.[20] If Armstrong replies that *p* would not have had a truthmaker, he would be stuck with the surely unwanted conclusion that it is possible for a proposition to be true without a truthmaker. And so the proper Armstrongian conclusion is that *p* would have had a different set of truthmakers, had there been no language-users. But then we are left with the question of what these truthmakers would have been—and until this question is answered, Armstrong's account is incomplete.

[18] Armstrong wrote: "There may be something mechanical, something purely conceptual, purely semantic, in the deductive following-out of proofs of the existence of the possible. (See the account of analytic truth to come in 8.9.)" (T&T, 102). Note that this quotation is from the chapter on *necessary* truths in T&T. So I take it that what Armstrong is (tentatively) suggesting here is that truths in mathematics about what is necessary are analytic.

[19] When Armstrong says that "a traditional view" is that analytic truths are "true solely in virtue of the meanings of the terms in which they are expressed," his wording seems to derive from the introduction to Ayer (1946). I think that few philosophers of mathematics today would defend Ayer's view in all its details. However, there are still philosophers who endorse views which resemble Ayer's position in important respects. See for example Rayo (2013).

[20] Perhaps some philosophers will insist that words (or their meanings) are necessarily existing abstract objects. However, I take it that *Armstrong* would not take this line. As we've seen, Armstrong rejected necessary abstracta.

6 Armstrong on Truthmaking and Necessity (Part 2: Sketch for a Systematic Metaphysics)

By the time he wrote *Sketch for a Systematic Metaphysics,* Armstrong had decided to reject his earlier suggestion that mathematical truths are analytic, saying that such a view implies that mathematics is "too arbitrary or conventional" (SSM, 91). But he suggested an alternative approach, which we will now consider.

We should begin by looking at Armstrong's metaphysics of law. Armstrong claimed that a law is a relation between properties (SSM, 35). Here is a toy example. Suppose that it is a law that being dehydrated causes headaches. For Armstrong, this means that a certain relation (viz. $\mathcal{N}$, the nomic relation) obtains between two properties (viz. the property *is dehydrated*, and the property *has a headache*). Armstrong would symbolize this as follows:[21]

$$\mathcal{N}(\textit{is dehydrated}, \textit{has a headache})$$

Now Armstrong claimed that laws have "instantiations." Our law, for example, is instantiated whenever someone is dehydrated and, consequently, has a headache. A law, on this view, is itself a property. And we have already seen that Armstrong was happy to posit "immanent" properties. This led Armstrong to the view that every law is instantiated. He wrote:

> If laws are a species of universal, then, according to me at least, they have to be instantiated at some place and time. Well, we talk of laws being instantiated, do we not? (The points where the laws are 'operative'.) So this instantiation of laws is the instantiation of a special sort of universal. (Note that this would require every law to be somewhere instantiated in space-time.) [...] One consequence of this is that there cannot be laws that are never instantiated. (SSM, 41)

Now Armstrong suggested that it is certain mathematical and logical laws which make true the necessities of mathematics.

He began his discussion of this proposal by appealing to his ENTAILMENT PRINCIPLE, arguing that truthmakers for the axioms of a mathematical theory

[21] I allow myself here to omit some of the finer details of Armstrong's account—in particular, I do not mention "state of affairs types" (SSM, 28–40).

doi: 10.48106/dial.v74.i4.05

must also be truthmakers for the theorems (SSM, 90). This manoeuvre is suspect. While it is known that the theorems of orthodox mathematics are entailed by the axioms (that's what makes them theorems, after all) it is far from clear that the axioms entail* the theorems.[22]

And it gets worse. To complete his account, Armstrong still needed to specify truthmakers for the axioms of our mathematical theories. To do this, he appealed to his theory of laws:

> We do, of course, have to recognize that introducing the Entailment Principle drives us back to consider the axioms from which mathematical systems are developed. [...] I suggest that we should postulate laws in logic and mathematics (non-contradiction, excluded middle in logic, Peano's axioms for number, or whatever laws logicians and mathematicians wish to postulate). In the light of the nature of proof just argued for we might suggest that such laws might be all we needed to postulate in the way of an ontology for logical and mathematical entities. (SSM, 90–91)

It is not credible, as Armstrong suggests here, that the Peano axioms are "laws" in Armstrong's sense. For example, one of the Peano axioms states that the natural numbers are *unending* in the sense that every natural number has a successor.[23] For Armstrong, this statement may not be true when taken at face value. For Armstrong, as we've seen, the existence of very large natural numbers is doubtful, and it is at least possible that there is a *largest* natural number, which has no successor. To circumvent this point, Armstrong will presumably insist on a modal reinterpretation of the axiom. On this view, the axiom, properly interpreted, states that, necessarily, every model of the Peano axioms is unending.

The corresponding Armstrongian law would then have to be:

[22] Suppose, for example, that Linnebo (2022) is correct and entailment* coincides with intuitionistic entailment. Then consider some statement φ which is provable classically but not intuitionistically from the prevailing axioms. (For example, φ might be $(\psi \vee \neg\psi)$, where ψ is some statement independent of the prevailing axioms.) Assuming, with Armstrong, that the inferences of classical logic are all truth-preserving, and the axioms of orthodox mathematics are true, we can conclude that φ is true. However, because it is not entailed* by the prevailing axioms, we cannot identify a truthmaker for it using the proposed method.

[23] The "successor" of a natural number is the number that comes *immediately after it,* when the natural numbers are arranged in the customary fashion. So for example the successor of nineteen is twenty.

$\mathcal{N}$(*is a model of Peano arithmetic*, *is unending*)

But this is problematic, because Armstrong believed that all laws are instantiated in the physical world—and it is far from clear that this law is instantiated. It may be that the physical world is finite. However, every model of Peano arithmetic is infinite. And so it may be that there are no physical models of Peano arithmetic, in which case the above-mentioned law is uninstantiated.

We might be able to avoid this problem by arguing on empirical grounds that there are infinitely many physical objects. For example, we might appeal to the common (though admittedly contested) assumption in physics that space is infinitely divisible. However, the problem that I have just described will reassert itself when we turn our attention from Peano arithmetic to other branches of mathematics which posit a greater number of entities. The most extreme case is set theory. Any model of second-order ZFC would have to have a truly vast domain, containing $\beth_\omega$ elements and more. There is no empirical reason to think that there exist that many physical objects. So we are left with the conclusion that the Armstrongian laws corresponding to the axioms of set theory are uninstantiated.

To avoid these problems, an Armstrongian would have to list a number of basic principles for mathematics which express laws that *are* instantiated in the physical world, and argue that they entail* the truths of mathematics. I don't know that this is impossible, but it is far from obvious that it can be done. And even if it *could* be done, the problem of identifying truthmakers for facts about what is possible would remain.

7 An Alternative Approach

Let's review. Armstrong believed that mathematical entities are located within the physical world. For example, wherever there is a pair of things, there is the number two. However, Armstrong realized that the physical world is not large enough to accommodate all the entities posited by modern pure mathematics. So he adopted a modal interpretation of mathematics. For Armstrong, pure mathematics tells us not about what is, but about what could be and what must be. However, Armstrong believed that every truth has a truthmaker within the physical world, and so he was left with the unenviable task of identifying truthmakers for modal truths within the physical world. I have argued that he did not succeed.[24] In this final section, I would like to put forward an alternative approach—an approach which will, I hope, appeal to those impressed by Armstrong's metaphysical system.[25]

Armstrong accepted a version of the methodological principle known as Occam's razor. He rejected mathematical Platonism largely for this reason. A "Platonic realm of numbers," he wrote, is an "ontological extravagance" (T&T, 100).[26] However, Armstrong did not use his razor to excise supervenient entities. Supervenient entities, he thought, are an "ontological free lunch." For example, he did not think that universalism in mereology is objectionably unparsimonious:

> Whatever supervenes or, as we can also say, is entailed or necessitated, is not something ontologically additional to the subvenient, or necessitating, entity or entities. [...] The terminology of "nothing over and above" seems appropriate to the supervenient. [...] If the supervenient is not something ontologically additional, then this gives charter to, by exacting a low price for, an almost entirely permissive mereology. Do the number 42 and the Murrumbidgee River form a mereological whole? [...] The whole, if it exists, is certainly a strange and also an uninteresting object. But if it supervenes on its parts, and if as a consequence of supervening it is not something more than its parts, then there seems no objection to recognizing the whole. So in this essay permissive mereology, unrestricted mereological composition, is embraced. (1997, 12–13)

On an uncharitable interpretation of this passage, Armstrong's view was that if the existence of x necessitates the existence of y, then y is "nothing over and above" x. But this is hardly plausible. Perhaps God exists necessarily, but it would be grossly immodest for me to claim that God is nothing over and above me. Perhaps I could not have had different parents, in which case my existence necessitates theirs. But they would quite properly take exception to the suggestion that they are nothing ontologically additional to me.[27]

[24] It is worth noting in passing that Armstrong's theory of *propositions* was problematic in rather similar ways. On this point, see McDaniel (2005).

[25] For a very different approach, see Read (2010).

[26] Armstrong also had epistemological reasons for rejecting Platonism (SSM, 2). For lack of space, I do not discuss epistemology in this paper.

[27] For a more detailed discussion of these points, see Schulte (2014).

Cameron (2008) has suggested a more promising way of developing Armstrong's idea that supervenient entities are "free." To put it briefly, Cameron's proposal is as follows. Compare the following two propositions:

⟨m exists.⟩ (where m is a marriage, between Ashni and Ben)
⟨e exists.⟩ (where e is an electron)

The former proposition is made true by certain patterns of human activity—involving perhaps Ashni, Ben, a registrar, some pieces of paper, and some metal rings. Ashni and Ben's marriage is a *derivative* entity: its existence is explained by facts about things other than itself. The electron e is not derivative. The electron's existence is not explained by facts about other things; e itself is the only truthmaker for the proposition ⟨e exists⟩.[28]

More generally, Cameron's proposal is this. When x is fundamental, the only truthmaker for ⟨x exists⟩ is x itself. When x is derivative, ⟨x exists⟩ has a truthmaker other than x itself.[29]

Cameron adds that it is derivative entities *in this sense* that are an "ontological free lunch," to use Armstrong's phrase. In effect, Cameron replaces the familiar slogan "Do not multiply entities beyond necessity" with a variant: "Do not multiply *fundamental* entities beyond necessity". Since mereological compounds are non-fundamental, Cameron infers, mereological universalism is not objectionable on grounds of parsimony.[30]

Cameron briefly suggests an application of this idea to impure set theory. He proposes that an impure set is "nothing over and above" its elements, so there is no objection on grounds of parsimony to positing all those impure sets that can be built up from *basic elements* whose existence can already be established. On this view, there is no need to re-interpret set theory in a "possibilist" manner. We maintain that all the sets posited by set theorists really

[28] Sharp-eyed readers will note that in this section I assume an *explanatory* conception of truthmaking, according to which, when T is a truthmaker for p, T *explains the truth of* p. For discussion, see Cameron (2018).

[29] I have actually modified Cameron's proposal in a small way. Cameron's view is that when x is derivative, x is *not* a truthmaker for ⟨x exists⟩. I find this claim puzzling (How could x fail to make true ⟨x exists⟩?) and since it is inessential to my argument, I omit it.

[30] Suppose that a and b are fundamental objects, and that $(a + b)$ is their mereological sum. According to Cameron a and b *collectively* make true ⟨$(a + b)$ exists⟩. For Cameron, this proposition has no *single* truthmaker; rather there are some things which *together* make the proposition true. This is a subtlety of Cameron's view which I ignore in the main text, for simplicity.

doi: 10.48106/dial.v74.i4.05

do exist, although they are not fundamental. Let's develop this Cameronian proposal in more detail.

Why does the set {Jill, Joe} exist? I suggest that it exists because Jill exists, and because Joe exists—and that is all. Nothing more is needed. And so, I suggest, any truthmaker for ⟨Jill exists⟩ and ⟨Joe exists⟩ will also be a truthmaker for ⟨A exists⟩, where A is {Jill, Joe}. More generally:

(1) If T is a truthmaker for ⟨x exists⟩, for each x in a non-empty set X, then T is a truthmaker for ⟨X exists⟩ also.[31]

So much for propositions about the existence of sets. But a complete truthmaker theoretic account of the sets will also include an account of what the truthmakers are for other propositions, including propositions about the identity and distinctness of sets, and propositions about what is an element of what.

Let's start with identity. Suppose that someone asks us why Joe is identical to Joe—that is, we are asked why Joe is identical with himself. This is a very peculiar question. The best answer to it that I can come up with goes like this. For Joe to bear the identity relation to himself, it suffices that he exists. Self-identicality is not some additional characteristic that requires further explanation. Joe exists, and so he is self-identical. And that is that. If this is right, I suggest, any truthmaker for the proposition ⟨Joe exists⟩ must also be a truthmaker for ⟨Joe = Joe⟩. In general, any truthmaker for ⟨x exists⟩ will also be a truthmaker for ⟨$x = x$⟩.[32]

Something similar is plausible in the case of non-identity. If, bizarrely, we are asked why it is that Jill is not identical with Joe—if we are asked why they are two people and not one—all we can say in reply is that to be non-identical

[31] What about the empty set? One might be tempted to avoid the problem by denying that the empty set exists. This proposal, as Hazen (1991) argues, is less radical than it might first seem, and Armstrong did in some places express scepticism about the empty set (T&T, 114). However, given Armstrong's usual hostility towards philosophically motivated reforms to standard mathematical practice, I think it desirable, from an Armstrongian point of view, to preserve the empty set. So here is an alternative, inspired by Kit Fine's well-known discussion of zero-grounding (2012).

Armstrong generally supposed that a truthmaker will always be a *single thing*. But we might want to allow that a proposition can be made true by two things acting in concert, or three things, or four things, or more. For example, we might say that a, b and c collectively make true ⟨$\{a, b, c\}$ exists⟩. Taking this line of thought still further, we could argue that in some unusual cases a proposition is made true by *zero* things; as we might put it, such propositions are *trivially* made true. On this view, we may say that ⟨$\varnothing$ exists⟩ is trivially made true.

[32] For more detailed discussion of the question of how truths of identity are to be explained, see Burgess, A. (2012) and Shumener (2017).

the two Bidens need only exist. Jill exists. She is one person. Joe exists too. He is another. And that is all. There is nothing extra that Jill and Joe need to do or to be in order to be distinct—existing is enough. And so, I suggest, any truthmaker for ⟨Jill exists⟩ and ⟨Joe exists⟩ is a truthmaker also for ⟨Jill $\neq$ Joe⟩. More generally, if x and y exist and are distinct, any truthmaker for ⟨x exists⟩ and ⟨y exists⟩ must also be a truthmaker for ⟨$x \neq y$⟩.

I want to recommend a similar treatment of the relations of membership and non-membership. If we are asked why Joe is an element of his singleton, there is nothing we can say except that, for this to be so, it suffices that Joe and his singleton exist. No more is needed. And if we are asked why Joe is not an element of {Jill}, we can say only that it is enough that Joe and {Jill} exist. More generally, I suggest, if x is an element of Y, then any truthmaker for ⟨x exists⟩ and ⟨Y exists⟩ is a truthmaker too for ⟨$x \in$ Y⟩. And if x is not an element of Y, though they both exist, any truthmaker for ⟨x exists⟩ and ⟨Y exists⟩ is a truthmaker too for ⟨$x \notin$ Y⟩.

Let me put all of this in a rather different way. Let's say that a relation R is "strongly internal" if and only if the following condition is met: Necessarily, for any a and b, if a bears R to b then (1) a bears R to b at any world at which a and b both exist, and (2) at every such world, any truthmaker for ⟨a exists⟩ and ⟨b exists⟩ is also a truthmaker for ⟨a bears R to b⟩.[33] If R is a strongly internal relation and a bears R to b, then no explanation for this is required, beyond whatever is needed to account for the fact that the relata exist. My proposal is that the relations of identity, non-identity, membership and non-membership are strongly internal in this particular sense.[34] In summary:

(1) If T is a truthmaker for ⟨x exists⟩, for each x in a non-empty set X, then T is a truthmaker for ⟨X exists⟩ also.

[33] Armstrong said that a relation is internal if "given just the terms of the relation, the relation between them is necessitated" (T&T, 9). That is, given any relation R, R is internal (in Armstrong's sense) just in case the following is necessary: For any a and b, if a bears R to b, then at every world at which a and b exist, a bears R to b.

Clearly, any strongly internal relation is also internal in Armstrong's sense.

The converse, however, is open to dispute. Suppose *arguendo* that God exists necessarily. Then the relation *x and y are such that God exists* is internal, in Armstrong's sense. But it is doubtful that this relation is strongly internal, for it is hardly plausible that any truthmaker for ⟨Joe exists⟩ and ⟨Jill exists⟩ mustalso be a truthmaker for ⟨God exists⟩.

[34] Note that strongly internal relations need not be universals—they may be "second-rate" properties. In saying that non-membership is strongly internal, I do not assert that it is a genuine universal.

doi: 10.48106/dial.v74.i4.05

(2) If a relation R is "strongly internal," then whenever a bears R to b, any truthmaker for ⟨a exists⟩ and ⟨b exists⟩ is also a truthmaker for ⟨a bears R to b⟩.
(3) The relations of identity, non-identity, membership and non-membership are strongly internal.

Let's take this further. Suppose for example that T is a truthmaker for the proposition ⟨Jill exists⟩. Then by (1), T is also a truthmaker for each of these propositions:

$$\begin{aligned}&\langle \mathrm{S}_1 \text{ exists}\rangle, \text{ where } \mathrm{S}_1 = \{\text{Jill}\}\\ &\langle \mathrm{S}_2 \text{ exists}\rangle, \text{ where } \mathrm{S}_2 = \{\{\text{Jill}\}\}\\ &\langle \mathrm{S}_3 \text{ exists}\rangle, \text{ where } \mathrm{S}_3 = \{\{\{\text{Jill}\}\}\}\\ &\vdots\end{aligned}$$

By (1) again, T is also a truthmaker for ⟨S_ω exists⟩, where S_ω is the set $\{\text{Jill}, \mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3, \ldots\}$.

By (2) and (3), T is a truthmaker too for various propositions about the relations among these sets, propositions like $\langle \text{Jill} \neq \mathrm{S}_1\rangle$, $\langle \mathrm{S}_1 = \mathrm{S}_1\rangle$, $\langle \mathrm{S}_1 \neq \mathrm{S}_\omega\rangle$, $\langle \mathrm{S}_\omega = \mathrm{S}_\omega\rangle$, $\langle \text{Jill} \in \mathrm{S}_1\rangle$, $\langle \mathrm{S}_1 \in \mathrm{S}_\omega\rangle$, and $\langle \mathrm{S}_1 \notin \text{Jill}\rangle$.

We can go further still, into the uncountable. Given our account, T will be a truthmaker for ⟨S^* exists⟩, where S^* is the set of non-empty subsets of S_ω. S^* is an uncountable set. And T will be a truthmaker for ⟨S^{**} exists⟩, where S^{**} is the set of non-empty subsets of S^*—a set even larger than S^*. And proceeding in this way, we can locate in the physical world truthmakers for propositions concerning sets at all levels of the vertiginous set-theoretic hierarchy, including sets of arbitrarily high cardinality.

And what of Armstrong's claim that all entities exist "somewhere, somewhen" (SSM, 15)? Well, some readers may find it edifying to insist that a set is located wherever its elements are.[35] On this view, you are co-located with your singleton, and its singleton, and *its* singleton, and so on *ad infinitum.* I offer no objection to this proposal. But I find it hard to see how to *justify* the claim that sets have spatial locations, and more importantly it seems to me that we need not endorse this claim to earn the title "naturalist." We insist

[35] Maddy (1990) defends this view. I lack the space for a thorough treatment of Maddy's approach, but I would like to note in passing that Maddy's version involves *reforms* to standard set theory: to be specific, Maddy identifies individuals with their singletons (e.g., for Maddy, Socrates = {Socrates}) and she eschews pure sets. The Armstrongian approach that I recommend preserves set theory in its usual form. (On the issue of pure sets, see footnote 31).

that all fundamental objects are physical, and that all truths have physical truthmakers—and this is naturalism enough.

Back to the cardinal numbers. According to the current proposal, even if the fundamental objects are rather few, nevertheless the sets are fantastically numerous. This allows us to maintain Armstrong's original account of cardinal number without having to worry about the problem of size, and without recourse to possibilism. Given the current proposal, for example, $\beth_\omega$ *is* instantiated in the hierarchy of sets, even if there are only finitely many fundamental entities. If we add that it is not possible for there to be nothing,[36] we are left with the conclusion that the cardinal numbers exist necessarily.

Of course, a thorough truthmaker-theoretic account of mathematics would also cover functions, complex numbers, matrices, ordinal numbers, graphs, and all the other mathematical creatures. You will probably be relieved to hear that I don't intend to deal with all these topics now. It's time for a cup of tea, after all. But I hope that my discussion of sets and cardinals is sufficient to motivate cautious optimism about Armstrongian naturalism—despite the errors of detail that we have identified in Armstrong's discussions of the metaphysics of mathematics.[*]

Thomas M.E. Donaldson
0000-0001-6497-7038
Simon Fraser University, Canada
tmdonald@sfu.ca

References

Armstrong, David M. 1989. *A Combinatorial Theory of Possibility*. Cambridge: Cambridge University Press.

—. 1997. *A World of States of Affairs*. Cambridge: Cambridge University Press.

—. 2004. *Truth and Truthmakers*. Cambridge Studies in Philosophy. Cambridge: Cambridge University Press, doi:10.1017/CBO9780511487552.

—. 2007. "Truthmakers for Negative Truths and Truths of Mere Possibility." in *Metaphysics and Truthmakers*, edited by Jean-Maurice Monnoyer, pp. 99–104.

[36] Armstrong changed his mind on the question whether it is possible for there to be nothing. In Armstrong (1989, chap. 4, section IV) he claims that this is possible, but in T&T (105) he retracts the claim.

[*] I would like to thank two reviewers at *Dialectica* for their hard work, and for their thorough and enormously helpful comments. I would also like to thank Jennifer Wang, Bahar Raiszadeh, Haktan Akcin, and Dans Waxman and Marshall for comments on drafts. (I've been working on this paper on and off for a long time; I apologize if I'm missing some acknowledgements!)

Philosophische Analyse / Philosophical Analysis n. 18. Heusenstamm b. Frankfurt: Ontos Verlag, doi:10.1515/9783110326918.

—. 2010. *Sketch for a Systematic Metaphysics*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199590612.001.0001.

AYER, Alfred Jules. 1946. *Language, Truth and Logic*. 2nd ed. London: Victor Gollancz.

BARNES, Jonathan. 1985. "Aristotelian Arithmetic." *Revue de philosophie ancienne* 3: 97–133. Reprinted, in revised form, in Barnes (2011, 334–363).

—. 2011. *Method and Metaphysics*. Essays in Ancient Philosophy n. 1. Oxford: Oxford University Press. Edited by Maddalena Bonelli.

BERRY, Sharon E. 2018. "Modal Structuralism Simplified." *Canadian Journal of Philosophy* 48(2): 200–222, doi:10.1080/00455091.2017.1344502.

BURGESS, Alexis. 2012. "A Puzzle about Identity." *Thought* 1(2): 90–99, doi:10.1002/tht3.14.

BURGESS, John P. and ROSEN, Gideon. 1997. *A Subject With No Object: Strategies for Nominalistic Interpretations of Mathematics*. Oxford: Oxford University Press, doi:10.1093/0198250126.001.0001.

CAMERON, Ross P. 2008. "Truthmakers and Ontological Commitment: or, How To Deal with Complex Objects and Mathematical Ontology Without Getting into Trouble." *Philosophical Studies* 140(1): 1–18, doi:10.1007/s11098-008-9223-3.

—. 2018. "Truthmakers." in *The Oxford Handbook of Truth*, edited by Michael GLANZBERG, pp. 333–354. Oxford Handbooks. New York: Oxford University Press, doi:10.1093/oxfordhb/9780199557929.001.0001.

FINE, Kit. 2012. "Guide to Ground." in *Metaphysical Grounding. Understanding the Structure of Reality*, edited by Fabrice CORREIA and Benjamin Sebastian SCHNIEDER, pp. 37–80. Cambridge: Cambridge University Press, doi:10.1017/CBO9781139149136.002.

—. 2017. "Truthmaker Semantics." in *A Companion to the Philosophy of Language*, edited by Bob HALE, Crispin WRIGHT, and Alexander MILLER, 2nd ed., pp. 556–577. Blackwell Companions to Philosophy. Oxford: Basil Blackwell Publishers. First edition: Hale and Wright (1997), doi:10.1002/9781118972090.

FOX, John F. 1987. "Truthmaker." *Australasian Journal of Philosophy* 65(2): 188–207, doi:10.1080/00048408712342871.

HALE, Bob and WRIGHT, Crispin, eds. 1997. *A Companion to the Philosophy of Language*. Blackwell Companions to Philosophy. Oxford: Basil Blackwell Publishers. Second edition: Hale, Wright and Miller (2017).

HALE, Bob, WRIGHT, Crispin and MILLER, Alexander, eds. 2017. *A Companion to the Philosophy of Language*. 2nd ed. Blackwell Companions to Philosophy. Oxford: Basil Blackwell Publishers. First edition: Hale and Wright (1997), doi:10.1002/9781118972090.

HAZEN, Allen Patterson. 1991. "Small Sets." *Philosophical Studies* 63(1): 119–123, doi:10.1007%2FBF00376001.

HELLMAN, Geoffrey. 1989. *Mathematics Without Numbers: Towards a Modal-Structural Interpretation*. 11th ed. Oxford: Oxford University Press, doi:10.1093/0198240341.001.0001.

—. 1998. "Maoist Mathematics?" *Philosophia Mathematica* 6(3): 334–345, doi:10.1093/philmat/6.3.334.

KESSLER, Glenn. 1980. "Frege, Mill and the Foundations of Arithmetic." *The Journal of Philosophy* 77(2): 65–79, doi:10.2307/2025431.

LINNEBO, Øystein. 2022. "Generality Explained." *The Journal of Philosophy* 119(7): 349–379, doi:10.5840/jphil2022119725.

LÓPEZ DE SA, Dan. 2009. "Disjunctions, Conjunctions, and their Truthmakers." *Mind* 118(470): 417–425, doi:10.1093/mind/fzp063.

MADDY, Penelope. 1990. "Physicalist Platonism." in *Physicalism in Mathematics*, edited by Andrew David IRVINE, pp. 259–289. The University of Western Ontario Series in Philosophy of Science n. 45. Dordrecht: Kluwer Academic Publishers.

MCDANIEL, Kris. 2005. "Review of Armstrong (2004)." *Notre Dame Philosophical Reviews*, https://ndpr.nd.edu/reviews/truth-and-truthmakers/.

PAWL, Timothy. 2010. "The Possibility Principle and the Truthmakers for Modal Truths." *Australasian Journal of Philosophy* 88(3): 417–428, doi:10.1080/00048400903193353.

RAYO, Agustín. 2013. *The Construction of Logical Space*. Oxford: Oxford University Press, doi:10.1093/acprof:oso/9780199662623.001.0001.

READ, Stephen. 2010. "Necessary Truth and Proof." *Kriterion: Revista de Filosofia* 51(121): 47–67, doi:10.1590/S0100-512X2010000100003.

RESTALL, Greg. 1996. "Truthmakers, Entailment and Necessity." *Australasian Journal of Philosophy* 74(2): 331–340, doi:10.1080/00048409612347331.

RODRIGUEZ-PEREYRA, Gonzalo. 2006. "Truthmaking, Entailment, and the Conjunction Thesis." *Mind* 115(460): 957–982, doi:10.1093/mind/fzl957.

ROSEN, Gideon. 1995. "Armstrong on Classes as States of Affairs." *Australasian Journal of Philosophy* 73(4): 613–625, doi:10.1080/00048409512346971.

SCHULTE, Peter. 2014. "Can Truthmaker Theorists Claim Ontological Free Lunches?" *European Journal of Philosophy* 22(2): 249–268, doi:10.1111/j.1468-0378.2011.00491.x.

SHUMENER, Erica. 2017. "The Metaphysics of Identity: Is Identity Fundamental?" *Philosophy Compass* 12(1), doi:10.1111/phc3.e12397.

SIMONS, Peter M. 1982. "Against the Aggregate Theory of Number." *The Journal of Philosophy* 79(3): 163–167, doi:10.2307/2026072.

WANG, Jennifer. 2013. "From Combinatorialism to Primitivism." *Australasian Journal of Philosophy* 91(3): 535–554, doi:10.1080/00048402.2012.722114.

YARNELLE, John Edward. 1964. *An Introduction to Transfinite Mathematics*. Boston, Massachusetts: DC Heath; Company.

Published by *Philosophie.ch*

Verein philosophie.ch
Fabrikgässli 1
2502 Biel/Bienne
Switzerland
dialectica@philosophie.ch

https://dialectica.philosophie.ch/

ISSN 1746-8361

ISBN 1234-5678

Dialectica is supported by the Swiss Academy of Humanities and Social Sciences.

Abstracting and Indexing Services

The journal is indexed by the Arts and Humanities Citation Index, Current Contents, Current Mathematical Publications, Dietrich's Index Philosophicus, IBZ — Internationale Bibliographie der Geistes- und Sozialwissenschaftlichen Zeitschriftenliteratur, Internationale Bibliographie der Rezensionen Geistes- und Sozialwissenschaftlicher Literatur, Linguistics and Language Behavior Abstracts, Mathematical Reviews, MathSciNet, Periodicals Contents Index, Philosopher's Index, Repertoire Bibliographique de la Philosophie, Russian Academy of Sciences Bibliographies.

Contents