No doubt, some choices need to be made. But what accounts for the choices made here? Is it anything other than strong feeling and unreflective opinion?[10]

University of Miami, Coral Gables, FL, USA
bxb475@miami.edu

References

Cappelen, H. and E. Lepore. 2008. *Insensitive Semantics: A Defense of Semantic Minimalism and Speech Act Pluralism*. Hoboken: John Wiley & Sons.

Clarke, D. 2013. *The Equality of the Sexes: Three Feminist Texts of the Seventeenth Century*. Oxford: Oxford University Press.

Ganeri, J. 2001. *Philosophy in Classical India: The Proper Work of Reason*. London: Routledge.

Lenz, M. 2019. *How Not to Make Progress in Philosophy: A Note on Scott Soames's New Book*. Handling Ideas December 28, 2019. http://handlingideas.blog/2019/12/28/how-philosophy-does-not-make-progress-a-note-on-scott-soames-new-book/, last accessed 14 August 2020.

Matilal, B.K. 1990. *Logic, Language, and Reality: Indian Philosophy and Contemporary Issues*. Delhi: Motilal Banarsidass.

Saul, J.M. 2012. *Lying, Misleading, and What Is Said: An Exploration in Philosophy of Language and in Ethics*. Oxford: Oxford University Press.

Soames, S. 2010. *What Is Meaning*? Princeton: Princeton University Press.

Soames, S. 2013. Deferentialism: a post-originalist theory of legal interpretation. *Fordham Law Review* 82: 597–617.

Stanley, J. 2000. Context and logical form. *Linguistics and Philosophy* 23: 391–434.

10 I am grateful to Magdalena Balcerak Jackson for her thoughtful feedback on earlier drafts of this review.

Substantive Radical Interpretation and the Problem of Underdetermination

ANANDI HATTIANGADI

1. *Introduction*

Consider the situation of a radical interpreter, an ideal being who sets out to interpret an arbitrary agent, such as Sally. From the set of all logically possible interpretations – understood to be mappings from representational vehicles to meanings or contents – the radical interpreter must select the unique interpretation that correctly maps Sally's brain states and vocalizations to what she perceives, believes, desires and means (11).[1] What facts must the radical interpreter know in order to select the correct

1 *The Metaphysis of Representation*. By J. Robert G. Williams. Oxford University Press, 2020. 240 pp. Hereafter, all page references are to this book, unless otherwise specified.

Analysis Reviews Vol. 80 | Number 4 | October 2020 | pp. 822–833 doi: 10.1093/analys/anaa069

interpretation of Sally? The question is posed in epistemological terms, but is essentially metaphysical: what facts determine the semantic facts about the meanings and contents of Sally's representations? Reductionists hold that there is some class of non-semantic, non-intentional facts that determine the semantic and intentional facts, knowledge of which would suffice for the radical interpreter to arrive at the correct interpretation of Sally. Williams is a reductionist in this broad sense.[2] In *The Metaphysics of Representation*, he develops a majestic programme for the specification of the non-semantic, non-intentional facts that determine the semantic and the intentional.

Williams's book is both important and timely, raising anew the unresolved question whether the semantic and intentional facts are determined by the non-semantic, non-intentional facts. This question took centre stage in the second half of the 20th century, but the discussion gradually fizzled out, being replaced by a vague feeling among many that the semantic facts *could* be reductively explained, even if no one had been able to show *how*. Williams sets out to do just that. His most notable proposal is to ground representational content not just in the physical facts, but additionally in certain *normative* facts (13), and his sophisticated new account of the normativity of content lies at the core of his ambitious foundational account of the determination of the contents of perceptions, beliefs, desires and sentences of a natural language. The book is innovative, full of insight and richly rewarding to read. It promises to have a profound impact on the philosophical investigation into the foundations of representation.

I have no hope of doing justice to every element of this sweeping work. After giving a brief overview in §2, I will focus the critical discussion on Williams's account of the determination of the contents of attitudes and concepts, which lies at the heart of his proposal.[3] In §3, I will argue that, in many cases, Williams's account leaves it underdetermined which of several mutually incompatible interpretations correctly represents the contents of an agent's states of mind.

2. *Reduction in three layers*

Williams divides his account of the reduction of representation into three layers. In the most fundamental layer, what gets determined is what he calls 'source intentionality', the contents of perceptual experience and low-level motor intentions. Building on the work of Karen Neander (2017), Williams develops a teleoinformational account of the determination of these basic contents, that is, one that grounds basic contents in causal facts and biologically evolved functions. The contents of Sally's perceptual experiences constitute her evidence, while the contents of her low-level motor intentions constitute the space of possible actions from which she chooses. This provides input to the next layer up.

In the second layer, Williams builds on David Lewis's argument for the reduction of representational content based on the possibility of radical interpretation. Lewis's

2 Reductionism is often defined more narrowly, as the view that the semantic properties are *identical* to some class of non-semantic properties, or as the view that semantic concepts can be reductively *analysed*. Williams is not explicitly committed to reductionism in either of these senses.

3 In this discussion, concepts are understood to be mental entities, the constituents of thoughts. This is in line with Williams's understanding.

strategy is to argue that the physical facts determine the semantic and intentional facts by showing how it is possible for a radical interpreter, who starts out omniscient about the physical facts, to select the correct interpretation of an agent, without recourse to any semantic or intentional information. Lewis specifies a number of criteria for the choice of interpretation, such as, most notably, the *Principle of Charity* and the *Rationalization Principle* (Lewis 1974). The Principle of Charity concerns the ascription of contents to the agent's beliefs: they ought to be ascribed in such a way as to make her beliefs rational in light of her evidence according to some suitable inductive method (Lewis 1979: 534) such as Bayesian conditionalization (Lewis 1983: 374). The Rationalization Principle requires that we assign credence and value functions to an agent in such a way as to satisfy the axioms of decision theory and assign the highest expected value to the agent's behaviour (Lewis 1974: 336).

Williams proposes to depart from Lewis's account in several important ways. First, whereas Lewis takes the input to radical interpretation to be the physical facts, Williams expands the class of inputs to include the semantic facts determined in the first layer – concerning the contents of the agent's perceptual experiences and motor intentions. However, even with the addition of these semantic facts as input to radical interpretation, Williams argues that a Lewisian radical interpreter faces a problem of underdetermination. He shows that from the correct interpretation of Sally's beliefs and desires (Original), it is possible to construct a deviant interpretation (Paranoid), which coincides with Original on what attitudes Sally adopts towards a local spatio-temporal bubble surrounding her, but which disagrees on what goes on outside the bubble, for instance, by ascribing to Sally the belief that everything outside of the bubble is void. Original and Paranoid equally satisfy the relevant Lewisian criteria for choice of interpretation – the principles of Charity and Rationalization – leaving it indeterminate which of the two the radical interpreter should choose.

Williams's proposed solution to this problem is to add to Lewis's *structural* rationality criteria for interpretation choice a further *substantive* rationality criterion for interpretation choice, which tells the radical interpreter to make Sally out to be as responsive to reasons as possible (26). This involves adding to the base certain normative facts about reasons; specifically, facts about what Sally has normative reasons to believe and do. The thought is that though Original and Paranoid make Sally out to be equally structurally rational, Original makes her out to be more substantively rational, and thus is correct. Williams fills out this general account of the determination of the contents of the attitudes with more focused, 'local' accounts of the determination of a range of classes of concepts, including empirical inductive and explanatory concepts; quantifiers and logical concepts; and moral concepts, such as 'right' and 'wrong'. Each of these local accounts is interesting in its own right.

The contents of the agent's beliefs and desires, which are determined in the second layer, provide input to the third layer, in which the contents of sentences and utterances in a public language are determined. Here, Williams builds on Lewis's account of convention, which views agents as typically *truthful* in their assertions – saying what they believe – and *trusting* in the assertions of others – believing what they hear (134). Since the conventional account of meaning relies heavily on the contents of agents' attitudes, much of the work of the third layer has already been accomplished in the second. What the third layer gives rise to are further considerations that a radical interpreter should take into account in ascribing contents to concepts and attitudes, so as to make certain linguistic utterances substantively rational and in line with the conventional account of linguistic meaning.

This completes Williams's story of the determination of a core set of representations, all the way from the ground-level contents of perceptions to the linguistic meanings up above. Stepping back from the details, Williams's central claim is that the semantic and the intentional facts are determined by a class of base facts that are themselves neither semantic nor intentional, and he makes it clear that he takes determination to be a metaphysical relation, akin to grounding or constitution (xvii). That is, the ultimate aim is to give an account of *what it is* for an arbitrary representational vehicle to have a meaning or a content.

2.1 Underdetermination again

The claim that the semantic and intentional facts are determined by some class of base facts entails the *supervenience* of the semantic and intentional on the base. That is, if we let S be a statement of the semantic and intentional facts at our world, let B be a statement of all of the base facts and let T be a 'that's all' statement to the effect that nothing more exists than is needed to satisfy the sentence that precedes it, the claim that B determines S entails that:

Supervenience. It is metaphysically necessary that: if BT then S.[4]

Supervenience rules out the possibility of *deviant worlds* that are just like our world in base respects (and that's all), but that differ from our world in semantic or intentional respects. This sets a condition of adequacy on any account of the determination of representational content: any such account must entail the impossibility of all deviant worlds. In an interpretationist setting, this translates to a condition of adequacy on the method of interpretation choice: it must yield the correct interpretation and only the correct interpretation of an agent every time. If by using the designated method, the radical interpreter is left undecided between the correct interpretation and another, i, the possibility of a deviant world, in which everything is like our world in base respects but i is correct, is left open. Yet, this is exactly the predicament Williams's radical interpreter finds herself in when she attempts to interpret imperfectly substantively rational agents.

A *perfectly* substantively rational agent, Williams tells us, is one whose every belief, disbelief and suspension of belief is justified, and whose every act is morally right (26). An imperfectly rational agent is one who is not perfectly rational in every respect. Though many of us are imperfectly rational, according to Williams, the radical interpreter should nevertheless attempt to make every agent out to be as substantively rational as possible. His central claim is that,

> The correct interpretation of an agent x is the one which best rationalizes x's dispositions to act in the light of the courses of experience x undergoes. (98)

That is, even when interpreting an imperfect agent – even when interpreting an agent who is downright foolish or vicious – the radical interpreter should select that interpretation from the set of logically possible interpretations that maximizes her substantive rationality. This might seem initially puzzling, since an interpretation

4 The formulation follows Chalmers (2009), which builds on Lewis (1994).

that makes an imperfect agent out to be perfectly rational would obviously be incorrect. However, the aim is not to make all agents out to be perfectly rational, but to make them out to be as rational as possible in light of their experience. In the case of a foolish or vicious agent, making her out to be as rational as possible may still wind up making her out to be not very rational at all.

Nevertheless, imperfect agents pose a serious difficulty for substantive radical interpretation. In many cases, there is no unique interpretation that best rationalizes an imperfect agent's dispositions in light of her experience, so substantive radical interpretation leaves the choice of interpretation underdetermined. This difficulty derives from the fact that there are several different rationality-making features of an agent, such as coherence and responsiveness to reasons, as well as several different domains of reasons, such as epistemic, moral and prudential – all of which can pull in several different directions. In the interpretation of imperfectly rational agents, rationality in one respect often has to be traded off against irrationality in another; responsiveness to one kind of reason has to be traded off against unresponsiveness to another. As a result, it is often possible to find two interpretations i_i and i_j, which differ dramatically from one another in the beliefs and desires they ascribe to an agent, yet which make her out to be roughly equally rational, simply because i_i makes her rational in some respects at the expense of making her irrational in others, while the reverse is true of i_j. This difficulty crops up in relation to the local determination of several concepts, including the logical concepts of negation and conjunction. However, for simplicity, I will focus on the moral concepts here.

Williams holds that moral concepts are 'referentially stable'. That is, despite differences in environment, society and moral opinion, moral concepts concern a common subject-matter, so long as they play a certain conceptual role in relation to agents' non-cognitive attitudes, emotions and deliberation. The core of this thesis is stated schematically as follows (78).

> **Referential stability (schematic).** Necessarily, if an agent has a concept W that plays role R, then W denotes P.

Referential Stability (Schematic) applies generally to moral concepts, such as 'right', 'wrong', 'good', 'bad', 'obligatory' and 'permissible'. The detailed account of the determination of any particular moral concept depends on what turns out to be the correct answer to substantive, meta-ethical questions about its conceptual role , which needs to be plugged in for R, and substantive moral questions about what makes an action right or wrong, which settle what is to be plugged in for P. Though Williams is not committed to any particular meta-ethics or moral theory, he assumes for the sake of argument that our concept 'wrong' has a distinctive blame-centric role. That is, roughly, judging an act to be wrong leads one to blame, while judging an act not to be wrong prevents one from blaming; if W occupies this conceptual role, then W is the concept 'wrong'. Moreover, he assumes that the Kantians are right about the substantive matter of what wrongness consists in, and hence that P is the property of violating the categorical imperative.

The difficulty becomes apparent when we consider the challenge a radical interpreter faces in interpreting a convinced utilitarian, such as John Stuart Mill. Let's suppose that Mill has a concept W that plays the relevant blame-centric role in his cognitive life, and that figures in the moral beliefs he would express using the term 'wrong'. Mill, of course, disagrees with Kant. He believes that it is not the case that an act is W if and only if it violates the categorical imperative. Rather, he believes that an act is W if and only if it fails to maximize happiness. Moreover, in the course of his life, he considers many actual and hypothetical moral cases and forms moral beliefs about them. For instance, at a certain point, he considers a case in which arresting an innocent man to appease an angry mob will maximize happiness, and holds that, though arresting the innocent man violates the categorical imperative, it is not W. That is, he has the following non-normative belief about the innocent man case:

(1) Arresting the innocent man violates the categorical imperative, but maximizes happiness.

And he has the following normative beliefs involving the concept W:

(1) An act is W if and only if it fails to maximize happiness.
(2) It is not the case that an act is W if and only if it violates the categorical imperative.
(3) Arresting the innocent man violates the categorical imperative, but is not W.

Since the contents of non-normative concepts are not really up for grabs here, suppose that (1) has the following content (I use bold to indicate reference to a proposition):[5]

(1a) **Arresting the innocent man violates the categorical imperative, but maximizes happiness.**

The radical interpreter now sets out to interpret W. If she is guided by the referential stability thesis, together with the substantive assumptions we have made about R and P, the correct interpretation of Mill's concept W is that it picks out the property of violating the categorical imperative, since that is the only interpretation that renders Mill's dispositions to blame substantively rational. That is, if the radical interpreter adopts Williams's proposed strategy, she will arrive at interpretation i_1, which ascribes the following contents to beliefs (2)–(4):

(2a) **An act violates the categorical imperative if and only if it fails to maximize happiness.**

(3a) **It is not the case that an act violates the categorical imperative if and only if it violates the categorical imperative.**

(4a) **Arresting the innocent man violates the categorical imperative, but does not violate the categorical imperative.**

It is clear at a glance that i_1 ascribes incoherent contents to Mill's beliefs: (2a) is incompatible with (1a), which is a purely non-normative belief about the case involving the innocent man, while (3a) is the negation of a trivial truth and (4a) is inconsistent. Since Williams suggests that the radical interpreter should maximize rationality in the ascription of propositional contents to an agent's concepts and attitudes, he seems to assume that rationality pertains to propositional contents, as opposed, for instance, to

5 Williams is neutral with respect to the question whether propositions should be understood as structured entities or sets of possible worlds (cf. 103–7). How this question is resolved is not germane to the present discussion.

Fregean senses.[6] If rationality pertains to propositional contents, and it is irrational to have beliefs with incoherent propositional contents, i_1 renders Mill irrational in this respect. In addition, i_1 renders some of Mill's beliefs epistemically unjustified and thus substantively irrational. For instance, when Mill considers the case of the innocent man, he acquires the non-normative belief that **arresting the innocent man violates the categorical imperative**, on the basis of which he forms the normative belief that arresting the innocent man is not W. If i_1 maps W to the property of violating the categorical imperative, then it maps this belief to the proposition that **arresting the innocent man does not violate the categorical imperative**, which is not justified by his non-normative belief.

Now consider an alternative interpretation, i_2, which maps W to the property of maximizing happiness. This interpretation assigns the following propositional contents to beliefs (2)–(4):

(2b) **An act fails to maximize happiness if and only if it fails to maximize happiness.**

(3b) **It is not the case that an act fails to maximize happiness if and only if it violates the categorical imperative.**

(4b) **Arresting the innocent man violates the categorical imperative, but does not fail to maximize happiness.**

In contrast to i_1, i_2 at least makes the contents of Mill's beliefs (1)–(4) coherent, if somewhat uninteresting. Indeed, (2b) is trivial, while (3b), and (4b), are perfectly sensible, and all three are compatible with (1a). Of course, by mapping W to the property of failing to maximize happiness, i_2 makes Mill's dispositions to blame unreasonable, by Williams' lights, because if Mill is disposed to blame people for performing acts that fail to maximize happiness, he is disposed to blame them for the wrong reasons. However, this is arguably compensated for by the fact that i_2 makes the contents of Mill's beliefs coherent. On the face of it, i_2 appears to be roughly equally as good at maximizing Mill's substantive rationality overall as i_1. At any rate, without any explicit method of aggregating different respects of rationality and different sorts of reasons, the two interpretations seem to rationalize Mill equally well overall. Thus, substantive radical interpretation, together with the referential stability thesis, underdetermines the correct interpretation.

Williams's discussion of two contrasting agents suggests a potential response to the foregoing difficulty. First, he considers Sally, whose moral beliefs about particular cases are derived from her more fundamental general moral beliefs, and whose general moral beliefs are reasonable, since they were formed on the basis of considering a wide range of cases, and consulting kith and kin (86). On Williams's view, Sally is largely substantively rational, though some of her moral beliefs are false, and because she is largely substantively rational, an interpretation that maximizes her substantive rationality is correct. Now, Sally is contrasted with Suzy, who spontaneously forms her particular moral beliefs about cases, rather than deriving them from any general moral beliefs. Suzy, Williams claims, is wildly irrational, and her concept W is tonk-like (86).[7] But

6 I discuss a Fregean response to this difficulty below.

7 It is not obvious that Suzy's concept is defective. Her only crime is not to derive her particular moral judgements about cases from general moral principles, but to formulate them spontaneously. In other words, she is a *particularist*. It would be surprising if it turned out that the particularist's moral concepts defy interpretation.

since Suzy is such an unlikely character, she poses no serious difficulty for substantive radical interpretation, which works for ordinary agents. This suggests the following response to the underdetermination problem: Mill is more like Sally than like Suzy, since his particular moral belief (4), is derived from his general moral beliefs (2) and (3), which in turn are reasonable, since they were formed responsibly. Because Mill is largely substantively rational, an interpretation that maximizes his substantive rationality is correct.

This response misses the source of the difficulty, which stems from the *contents* that i_1 assigns to Mill's beliefs, not to the process by which Mill arrives at those beliefs. For, even if Mill were to derive (4) from (2) and (3), like Sally, his derivation would be substantively irrational if the contents of these beliefs are as ascribed by i_1. This is because the belief that **arresting the innocent man violates the categorical imperative, but does not violate the categorical imperative** is not justified by the belief that **an act violates the categorical imperative if and only if it fails to maximize happiness**, together with belief that **it is not the case that an act violates the categorical imperative if and only if it violates the categorical imperative.**[8] Moreover, even if Mill is like Sally in his later life, deriving his particular beliefs about cases from his general moral beliefs, at some point earlier in his life, when he encountered a moral case for the first time, he must have spontaneously formed a particular moral belief without deriving it from a general one. Indeed, if the general moral beliefs he arrives at in later life are justified by careful consideration of many cases, as Williams suggests, then Mill will have made many spontaneous moral judgements about many cases before arriving at his general moral belief. If we focus on the interpretation of Mill during his formative years, interpretation i_1 does not make him out to be more substantively rational than i_2. Thus, no details about the process by which Mill arrives at his moral beliefs addresses the fact that some of the contents i_1 ascribes to Mill's are straightforwardly incoherent.

One reaction to the foregoing, which will come naturally to Fregeans, is that the problem arises from the fact that reference and content are *non-transparent*. As Frege (1892/1997) famously observed, it is possible for a rational agent to believe that Hesperus is bright, and not believe that Phosphorus is bright, though Hesperus = Phosphorus. He postulated sense to account for the difference in cognitive significance between these two beliefs and argued that rationality pertains to sense, rather than reference or propositional content. On this view, it is possible for a rational agent to believe that Hesperus is not Phosphorus, though the propositional content of her belief is that **Venus is not Venus**, which is inconsistent.

Though Williams's official story is that the radical interpreter assigns propositional contents to beliefs in a way that maximizes an agent's substantive rationality, he expresses some sympathy for a Fregean approach and even goes so far as to defend an interpretationist account of the determination of sense. It is worth considering therefore whether the problem of underdetermination can be resolved if interpretations map beliefs to senses in addition to mapping them to their contents.

Williams's account of the determination of sense goes as follows. Building on Evans, he claims that the sense of an agent's general concept C is ***standing in relation R to property P***, where it is in virtue of an agent's standing in R to P that her concept C refers

8 Of course, since (3a) is a contradiction, (4a) follows from it by Explosion, which states that everything follows from a contradiction. Nevertheless, belief in a contradiction does not *justify* belief in what that follows from it.

to P. (Note that I use bold italics to indicate that I am referring to senses). He then argues that to determine the sense of a concept C, the radical interpreter need only work out the relation R in which the agent stands to P, and in virtue of which C has P as its content (107). According to the referential stability thesis, the content of a moral concept is determined by its conceptual role, which suggests that in the case of moral concepts, their sense adverts to their conceptual role. Given the substantive assumptions that we have made for the sake of argument, then, the sense of W can be roughly glossed as ***blameworthy***. With this in place, consider an interpretation i_3, which ascribes senses rather than propositional contents, and maps Mill's beliefs (1)–(4) to the following senses:

(1c) ***Arresting the innocent man violates the categorical imperative, but maximizes happiness.***

(2c) ***An act is blameworthy if and only if it fails to maximize happiness.***

(3c) ***It is not the case that an act is blameworthy if and only if it violates the categorical imperative.***

(4c) ***Arresting the innocent man violates the categorical imperative, but is not blameworthy.***

At first blush, this appears to be a promising solution to the problem, since the senses ascribed by i_3 to Mill's beliefs (1)–(4) are coherent.[9] However, i_3 nevertheless renders Mill's beliefs epistemically unjustified by his evidence. For instance, Mill's belief that ***arresting the innocent man is not blameworthy*** is based on his belief that ***arresting the innocent man maximizes happiness***, but is not justified by it. This is because, by hypothesis, the fact that an act maximizes happiness is not a substantive, normative reason to believe that it is not blameworthy. Since substantive rationality has to do with responsiveness to reasons, i_3 makes Mill out to be unresponsive to epistemic reasons in the formation of his moral beliefs. As a consequence, if the radical interpreter is to choose now between *pairs* of interpretations – one that maps attitudes to contents and the other that maps attitudes to senses – she is faced with two pairs of interpretations of Mill, $\langle i_1, i_3\rangle$ and $\langle i_2, i_3\rangle$, which rationalize Mill's dispositions in light of his experiences roughly equally well.[10]

9 Of course, Mill, like many utilitarians, dissociates rightness and wrongness from praise and blame, on the grounds that praising and blaming are themselves acts, which can be evaluated as right or wrong in themselves. Since utilitarians hold that an act is right if and only if it maximizes happiness, it is not necessarily the case that the act of praising someone for performing an act that maximizes happiness will itself maximize happiness, nor that blaming someone for performing an act that doesn't maximize happiness will maximize happiness. If Mill also believes that it is possible for an act to be wrong but not blameworthy, i_3 ascribes that belief the incoherent sense ***it is possible for an act to be blameworthy but not blameworthy***.

10 Williams (personal communication) suggests that on Evans's proposal, $\langle i_2, i_3\rangle$ will not be in good order because the senses ascribed in i_3 will not align with the properties ascribed in i_2. This is because on Evans's proposal it is supposed to be possible to 'read off' the referent of a concept from its sense. Applied to the case at hand, if the sense of W is ***blameworthy***, it follows that W picks out the property of violating the categorical imperative. However, this move is in tension with substantive radical interpretation, which states that the radical interpreter should maximize rationality overall. Adding a further constraint on the relation

Furthermore, the appeal to senses introduces a new source of indeterminacy. We have hypothesized that our concept 'wrong' has a certain blame-centric conceptual role. But there is another concept, which we can call 'wrong*', whose sense can be glossed as ***to be avoided***, and which has a motivation-centric conceptual role: roughly, judging an act to be wrong* leads one to refrain from performing that act, while judging an act not to be wrong* does not lead one to refrain from performing it. Now, let's suppose that Mill tends to refrain from performing acts he judges to be W and not to refrain from performing acts he judges to be not W. This is plausible, and compatible with his having the dispositions to blame that we have already assumed him to have. This leaves open the possibility that Mill's concept W is 'wrong*', and hence that the senses of (2)–(4) are as ascribed by i_4:

(2d) **An act is to be avoided if and only if it fails to maximize happiness.**

(3d) **It is not the case that an act is to be avoided if and only if it violates the categorical imperative.**

(4d) **Arresting the innocent man violates the categorical imperative, but is not to be avoided.**

Once again, i_4 makes Mill's beliefs coherent but renders his moral beliefs about what acts are to be avoided unjustified by his normative epistemic reasons. For instance, Mill's belief that ***arresting the innocent man is not to be avoided*** is based on his belief that ***arresting the innocent man maximizes happiness***, but is not justified by it. This is because, assuming that Kant is right, the fact that an act maximizes happiness is not a substantive, normative reason to believe that it is not to be avoided. Since substantive rationality has to do with responsiveness to reasons, i_4 makes Mill out to be unresponsive to epistemic reasons in the formation of his moral beliefs. Now, the radical interpreter is faced with *four* pairs of interpretations of Mill, $\langle i_1, i_3\rangle$, $\langle i_1, i_4\rangle$, $\langle i_2, i_3\rangle$ and $\langle i_2, i_4\rangle$, all of which rationalize Mill's dispositions in light of his experiences roughly equally well.

In the face of underdetermination of this kind, which arises from the fact that several interpretations make an agent out to be rational in some respects but irrational in others, Williams suggests that what the radical interpreter needs is a way of aggregating the information about how rational an interpretation renders an agent in several different respects into a measure of how rational it renders the agent overall. To dramatize, imagine that the radical interpreter is aided by a committee, each member of which cares only about how rational an interpretation makes the agent out to be in some particular respect: Ms Coherent cares only how coherent an interpretation makes the agent out to be, while Ms Morality cares only how far an interpretation goes in making the agent responsive to substantive moral reasons. Now, each committee member ranks the interpretations in the set of logically possible interpretations in line with their concerns; let's call any set of individual rankings a *profile*. What the radical interpreter needs is a voting procedure, an aggregation function, f, from any such profile of individual rankings to a ranking of interpretations overall, which selects the correct interpretation of an agent every time.

between the ascribed sense and the referent only adds a new source of potential indeterminacy. How is this fact about interpretations to be weighed against the extent to which they make an agent rational in other respects?

Though it is true that there *is* a function such as this, it is but one among *zillions* of functions from profiles of individual rankings to rankings of interpretations overall. How is the radical interpreter to select the needle in this haystack? What she needs are some constraints on the choice of an *aggregation function* that rule out all but the one that delivers the correct interpretation every time. I have argued elsewhere that the situation of the radical interpreter can be modelled with resources from social choice theory, and that a modified version of Arrow's theorem can be applied to show that there is no aggregation function that satisfies a small set of intuitively plausible constraints (Hattiangadi 2019). Though I focused on Lewis's criteria of interpretation choice, the point generalizes to Williams's proposal. To avoid the impossibility, the Arrovian constraints on aggregation must be replaced by another set of intuitively plausible constraints, which narrow the choice of function down to the one that delivers the correct interpretation of each and every agent. In the absence of some such account, the radical interpreter's choice of interpretation remains severely underdetermined.

3. *Conclusion*

The upshot of the foregoing is that Williams has not shown that the semantic and intentional facts supervene on the designated base facts, because the account of the determination of representational meaning and content does not rule out the possibility of deviant worlds. Suppose, for the sake of argument, that the correct interpretation of Mill is $\langle i_1, i_4\rangle$; that this interpretation maps Mill's mental states to the senses and contents that they actually have. Now consider a world w that is just like our world in all base respects, but at which Twin-Mill's states of mind are correctly represented by a different interpretation, $\langle i_2, i_3\rangle$. Since $\langle i_2, i_3\rangle$ makes Twin-Mill out to be just as rational as $\langle i_1, i_4\rangle$ makes Mill out to be, substantive radical interpretation does not rule out the possibility of w. But if w is possible, the semantic and intentional facts do not supervene on the base facts.

It is worth underscoring just how extensive this underdetermination is. Mill is not an outlier, with wildly irrational attitudes and tonk-like concepts. He is an ordinary utilitarian. If substantive radical interpretation leaves Mill's state of mind indeterminate, the same goes for virtue theorists and perfectionists, not to mention the morally depraved.[11] Indeed, this underdetermination generalizes to the local accounts of many different concepts, since in every case, the radical interpreter is encouraged to maximize substantive rationality, though rationality in one respect often has to be traded off against irrationality in another. Since most of us do not even approximate perfect substantive rationality, substantive radical interpretation leaves the states of mind of most ordinary people underdetermined.[12]

Stockholm University
10691 Stockholm, Sweden
anandi.hattiangadi@philosophy.su.se

11 This is against the background assumption that Kant was right, of course. If Mill was right, then radical interpretation leaves it indeterminate what attitudes all non-utilitarians hold. The point, of course, remains: underdetermination is widespread.

12 Thanks to Alexander Sandgren and Robert Williams for comments and discussion.

References

Chalmers, D. 2009. The two-dimensional argument against materialism. In *Oxford Handbook to the Philosophy of Mind*, eds. B.P. McLaughlin and S. Walter, 313–38. Oxford: Oxford University Press.

Frege, G. 1892/1997. On Sinn and Bedeutung. In *The Frege Reader*, ed. M. Beaney, 151–71. Oxford: Blackwell.

Hattiangadi, A. 2019. Radical interpertation and the aggregation problem. *Philosophy and Phenomenological Research*. doi:10.1111/phpr.12578.

Lewis, D. 1979. Attitudes De Dicto and De Se. *The Philosophical Review* 88: 513–43.

Lewis, D.K. 1974. Radical interpretation. *Synthese* 27: 331–44.

Lewis, D.K. 1983. New work for a theory of universals. *Australasian Journal of Philosophy* 61: 343–77.

Lewis, D.K. 1994. Reduction of mind. In *Companion to the Philosophy of Mind*, ed. S. Guttenplan, 412–31. Oxford: Blackwell.

Neander, K. 2017. *A Mark of the Mental: In Defense of Informational Teleosemantics*. Cambridge, MA: MIT Press.

Williams, J.R.G. 2020. *The Metaphysics of Representation*. Oxford: Oxford University Press.