

Gruesome Diagonals

Laura Schroeter

Philosophers' Imprint
<www.philosophersimprint.org/003003/>
Volume 3, No. 3
August 2003
©2003 *Laura Schroeter*

How much work can two-dimensional semantics do? An awful lot, according to Frank Jackson and David Chalmers. For starters, Jackson and Chalmers have argued that their 2-D framework allows us to reconcile semantic externalism with the traditional idea that subjects have a priori access to what it takes to fall into the extension of their own representations. You might have thought that the lesson of Twin Earth thought experiments was that subjects need not know such facts. But according to Jackson and Chalmers, this is the wrong moral to draw. You may not know that 'water' refers to H₂O, but you do know that *if* the actual world is like we think it is then water is H₂O, and *if* the actual world is like Putnam's Twin Earth then water is XYZ. In general, Jackson and Chalmers conclude, subjects can know a priori how the extension of their representations varies depending on what the actual world turns out to be like. The 2-D framework is simply meant to summarize the subject's ideally rational judgments about the extension of 'water' relative to hypotheses about what actual and counterfactual worlds might be like. The diagonal of this 2-D matrix captures what the subject can know a priori—roughly, that water is the watery stuff around here. What the subject can only know a posteriori—that water is H₂O—is characterized by the horizontal row corresponding to the actual world.

This claim about the a priori accessibility of what our words and thoughts represent is an epistemological thesis. But Jackson and Chalmers also use their 2-D framework to defend a genuinely semantic thesis: the diagonal intensions defined by the 2-D matrixes can be used to characterize the meanings of words and the contents of thoughts. According to common sense, the meaning of 'water' is not exhausted by the fact that it refers to H₂O—for the term 'H₂O' also refers to H₂O, and yet we think that the two terms have distinct meanings. Meanings and thought contents reflect the

Laura Schroeter is a Research Fellow at Monash University.

fine-grained representational states commonsense psychology appeals to in characterizing rational communication and cognition. Jackson's and Chalmers's suggestion is that diagonal intensions provide a perspicuous representation of the intuitively correct identity conditions for these representational states. In particular, Jackson has suggested, diagonal intensions capture the *conventional linguistic meaning* associated with public language expressions like 'water'. Chalmers, on the other hand, has focused more on the semantic properties of individual subjects' words and thoughts: diagonal intensions, he thinks, capture the *units of cognitive significance* expressed by a subject's words and thoughts.

If Jackson and Chalmers are right, diagonal intensions simultaneously play two key theoretical roles. On the epistemological side, diagonal intensions provide a priori accessible extension conditions for token representational states. On the semantic side, diagonal intensions individuate the fine-grained representational state types needed to characterize the rational evolution of thought and talk. If their project is successful, then Jackson and Chalmers will have vindicated the two most important ambitions of a Fregean theory of sense.

It is Jackson's and Chalmers's epistemological thesis which has attracted the most critical attention. In this paper, however, my target will be the semantic thesis. Even if Jackson and Chalmers are right that diagonal intensions can be used to characterize what subjects can know a priori about the extensions of their words and thoughts, it is still an open question whether semantic theorists should take diagonal intensions to characterize the meaning of words or the content of thoughts. I'll argue that diagonal intensions cannot play the semantic roles Jackson and Chalmers assign to them.

1. *Two models of 2-D semantics*

According to Jackson and Chalmers, 2-D matrixes are assigned to a subject's representations on the basis of her an-

swers to questions of the following form:

If the actual world turns out to be thus-and-so, which items count as water in this world?

If the actual world turns out to be thus-and-so, which things count as water in such-and-such counterfactual world?

The *diagonal intension* for the subject's word 'water' is determined by the answers she would give to the first type of question after idealized rational reflection on the empirical facts about the world considered as actual. The *horizontal intensions* are determined by her ideal answers to the second type of question. Since all the empirical information which is relevant to answering these diagnostic questions is built into the hypothesis about the actual world, the subject's answers will not depend for their justification on empirical knowledge of her real environment. Thus, on Jackson's and Chalmers's account, the 2-D matrix assigned to the subject's word 'water' is determined by the subject's own ideal a priori verdicts about each cell of the matrix.¹

In this section, I'll take a closer look at how a subject might go about answering Jackson's and Chalmers's diagnostic questions. My aim is to get a finer-grained picture of the kinds of judgments we make about hypothetical scenarios, and the kinds of considerations we take to justify those judgments. This will put us in a position to appreciate what sort of psychological abilities ground the assignment of

¹See Chalmers and Jackson, "Conceptual Analysis and Reductive Explanation" for an overview of their approach. In order to formulate the diagnostic questions, Jackson and Chalmers invoke an ideal vocabulary suitable for characterizing every epistemically possible way the world might be. Their suggestion is that this vocabulary might include terms for all possible microphysical and phenomenal properties, the indexical expressions 'I' and 'here', and a 'that's all' clause to indicate that the description is complete. I have criticized aspects of this general framework elsewhere. However, for present purposes, I want to set aside any qualms about the adequacy of Jackson's and Chalmers's 2-D framework.

a particular diagonal intension to a subject's representation. These abilities, I'll argue, are more complex than is often assumed. The exclusive focus on Twin Earth in the literature can be misleading in this respect. Twin Earth is an easy case, as our judgments about it tend to be immediate and unequivocal. Once we expand our diet of examples to include other hypotheses about what the actual world might turn out to be like, it becomes much more difficult to decide what falls into the extension of a word like 'water'. In order to see this, we'll need to run through some examples and test our intuitions about how to answer Jackson's and Chalmers's diagnostic questions.

For simplicity, I'll assume that you're convinced that your word 'water' actually refers rigidly to the chemical kind H_2O . That is, you're convinced that to be a sample of water *just is* to be a sample of H_2O : the two properties are co-extensive both in the actual world and in all counterfactual worlds.² If this is your starting point, then it will be a remarkably easy and straightforward matter to decide which things count as water on the hypothesis that the actual world turns out to be like Twin Earth. You have amassed a huge number of beliefs and cognitive dispositions regarding water: you're opinionated about water's perceptual gestalt (look, taste, odor), its physical characteristics (heaviness, viscosity, rough boiling point, solvent role), its ecological role (rain, rivers, erosion), its biological role (necessary for plants and animals), its role in practical life (agriculture, transportation, washing clothes, making tea, surfing), its explanatory role (a unified explanatory kind underlying the perceptual, physical, and biological characteristics), its epistemological

²Of course, not everyone agrees with this prevailing philosophical orthodoxy (and I count myself among the skeptics). The assumption, however, is harmless for my purposes. If you think that your word 'water' is ambiguous, or that it picks out a functional kind, or that it has some vagueness that doesn't match that of the scientific kind H_2O , then you'll probably already be aware of the hermeneutical complexities of trying to defend your preferred interpretation against competing ones.

accessibility (easy to spot, hard to define), as well as about its essential nature (a chemical compound, H_2O). The Twin Earth story leaves all of these tacit and explicit beliefs unchallenged except the very last one. When you consider Twin Earth as actual, you simply imagine that a different chemical compound, XYZ, has all of the important perceptual, physical, ecological, biological, practical, explanatory, and epistemological characteristics you currently associate with H_2O . It's even essentially definable in terms of chemical theory. So what counts as water, if the actual world turns out to be like that? If you currently think that water is H_2O , it's hard to see how you could doubt that the right answer is: XYZ.

Putnam's story is custom designed to make this decision easy. And we can invent a large range of cases that look just as straightforward. For instance, it's coherent to suppose that the stuff that plays all the various roles you currently associate with water is a certain configuration of Aristotelian prime matter. If that's what the actual world is like, then Aristotle was right: water is a kind of prime matter. Or the stuff that plays those roles might turn out to be a kind of living organism—something like a slime mold. Unlike the Twin Earth scenario, these scenarios falsify your current conviction that the underlying explanatory kind that plays all those roles is definable within the framework of modern chemistry. But in view of the overwhelming vindication of your other tacit and explicit beliefs about water, this seems unimportant. After all, it's central to your understanding of water that it's an explanatorily basic kind whose essence can only be known with precision on the basis of careful empirical inquiry.

When we focus on these Twin Earth-style scenarios, it's tempting to suppose that your understanding of the word 'water' might be structured in much the same way as your

understanding of a rigidified definite description. Perhaps 'water', as you understand it, might be glossed as 'the explanatory kind that actually plays the water-role around here', where 'the water-role' is shorthand for the total set of cognitive commitments you currently associate with the term 'water'. The key psychological assumption in this updated version of Fregean descriptivism is that the subject implicitly grasps an algorithm that settles the actual and counterfactual extension of her representation no matter what the actual world turns out to be like.³ According to the current proposal, you have a two-step algorithm that allows you to fill in the 2-D matrix for your term 'water'. First, your tacit understanding of the water-role and of what it is to be an explanatory kind allows you to determine, for every centered world considered as actual, which things count as water in that world. This allows you to fill in the diagonal intension of the matrix:⁴

³It's not essential to this new Fregean position that the subject be able to exhaustively define the relevant algorithm in natural language. After all, the subject's implicit understanding may depend essentially on pre-linguistic recognitional dispositions. The crucial assumption is that the subject have some stable, entrenched psychological dispositions which would allow her to identify the actual and counterfactual extensions of her representation mechanically when she is presented with the relevant empirical information. No substantive theoretical or practical deliberation is required, since the subject's standing dispositions determine precisely what it takes for something to fall into the actual and counterfactual extension of her representation.

⁴The vertical axis of the matrix represents all possible "centered" worlds (i.e. worlds with a designated time and subject) considered as the actual context of use; and the horizontal axis represents those same worlds, *without* a center considered as circumstances of evaluation. Chalmers has stressed that considering a world as actual should not be taken to imply that the subject or her thoughts exist in that world. My use of the phrases 'centered possible world considered as actual' and 'possible context of use considered as actual' is meant to be consistent with this presupposition. Unlike Chalmers, I will use these two phrases interchangeably. (Cf. Chalmers's "The Foundations of Two-Dimensional Semantics".)

| | | | |
|-----------------|--------|-------|--------|
| | W_1 | W_2 | $W...$ |
| ██████ W^*_1 | H_2O | | |
| ██████ W^*_2 | | XYZ | |
| ██████ $W^*...$ | | | ... |

With this information in hand, our understanding of the nature of explanatory kinds will put you in a position to determine which items in counterfactual worlds belong to the very same explanatory kind as samples of water in the centered world considered as actual. So you can then fill in the horizontal rows of the matrix:

| | | | |
|-----------------|--------|--------|-----------|
| | W_1 | W_2 | $W...$ |
| ██████ W^*_1 | H_2O | H_2O | $H_2O...$ |
| ██████ W^*_2 | XYZ | XYZ | XYZ ... |
| ██████ $W^*...$ | | | ... |

Let's call this the *simple model* of your psychological ability to fill a 2-D matrix for a word like 'water'. This simple model would explain your immediate and unhesitating judgments about Twin-Earth-style cases, since it posits a clear-cut algorithm that guides you in filling in each cell of the matrix.

However, the model cannot handle other sorts of cases. As soon as we depart even in small ways from Putnam's script, it becomes much more difficult to answer the diagnostic questions that define the 2-D matrix. To see this, let's start with a standard variant on the Twin Earth scenario. It might turn out that your immediate environment includes *both* H_2O and XYZ and you fail to distinguish the two. In

that case, which samples of liquid count as water? In the imagined scenario, there is no single chemical kind that plays the roles you currently take H_2O to play. If you understood your word 'water' to be synonymous with the rigidified description 'the explanatory kind which actually fills the water-role', you'd say that this is a scenario where there is no water. But clearly this is implausible. After all, you've been using the word 'water' all your life to classify things in your environment in a systematic and useful way: you use the word to classify stuff as suitable for drinking, for putting out bushfires, for washing your clothes, and so on. If you learn on the evening news tonight that the stuff you've been classifying as water all these years is actually composed of two distinct chemical kinds, are you going to conclude that everything you ever said or thought using the word 'water' is *false*? Surely that would be a weird reaction. According to common sense, your past claims about water were mostly true—you just need to repudiate any claims you might have made about water having a single, unified scientific essence. But what exactly should you take the truth of your water beliefs to hinge on, if the actual world turns out to contain both H_2O and XYZ? What counts as water in this scenario?

It's generally agreed that which kind of stuff you should identify as water depends on the details of your interaction with the two substances, H_2O and XYZ. It may be that you systematically confound the two substances in your everyday activities, in much the way that naive subjects confound nephrite and jadeite. In that case, it may make sense to interpret your word as representing members of a disjunctive kind, H_2O -or-XYZ. On the other hand, it may turn out that your interaction with the two substances is quite neatly segregated: H_2O is the stuff that is naturally found around here, but XYZ is the stuff just over the hill and you've never had occasion to confound the two. In that case, it may make best sense to say that your word is ambiguous and to assign two distinct semantic values. In Hartry Field's terminology, your

word *partially refers* to H_2O and it *partially refers* to XYZ.⁵ This is quite different from saying that it refers to the disjunctive kind H_2O -or-XYZ. Clearly, deciding which of these competing interpretations is correct will be neither easy nor automatic. Even identifying the relevant interpretive options requires a good deal of cognitive effort. And of course there will be many double-satisfier scenarios where the correct interpretation will be much harder to determine than in the relatively clear-cut cases I've highlighted here.⁶

Next, let's consider a more radical departure from the Twin Earth paradigm. It could turn out that there is no systematic explanatory kind underlying the superficial watery phenomena in your environment. The stuff you've habitually classified as water is hopelessly heterogeneous at the level of chemistry. For instance, it might turn out that there are a number of distinct chemical substances which, in various mixtures and proportions, play the roles we associate with 'water'. Some of these mixtures might be more suitable to some roles than others. Perhaps the mixture X+Y tends to collect in rivers and lakes, but is not potable unless mixed with Z. Because of small differences in the microphysical structure together with the filtration dynamics of the soil, the stuff in aquifers tends to be pure Z or U, both of which are potable. Rain, rivers, and oceans contain various proportions of all of these substances, though different microclimates sometimes have a higher proportion of one or another. All these various mixtures look and taste roughly as you'd expect water to. Which stuff counts as water in this sort of scenario?

Clearly there is no easy answer to this question. Once again, it's implausible to say that all your thought and talk

⁵See Hartry Field's "Theory Change and the Indeterminacy of Reference" for the notion of partial reference.

⁶Ned Block and Robert Stalnaker make a similar point in "Conceptual Analysis, Dualism, and the Explanatory Gap", p. 21.

about water in the past has been systematically false. So it seems you must identify some items in the scenario as making your claims about water true. But which ones? One natural answer is to say your word represents some functional kind based on your non-explanatory interests in the term. Consider air, for instance: from a scientifically naive perspective, it seems just as explanatorily basic as water, but we've discovered that there is no basic chemical kind underlying airy phenomena. Our current use of 'air' seems to be ambiguous between the functional kind, *breathable stuff*, the ecological kind, *atmosphere*, and the vague chemical kind, *mixture of nitrogen and oxygen*. It might have turned out that 'water' suffered a similar fate. Thus in the imagined scenario, it might make best sense to say that water is a nutritional kind, something like *basic drink* or an ecological kind like *river-stuff*.

Whether one or another (or both) of these interest-based functional kinds should be taken to determine the extension of your term 'water' will depend in part on the details of your interests, beliefs, and cognitive dispositions. If you're an illiterate and isolated desert-dweller, for instance, it won't make much sense to take the truth of your claims about water to turn on the ecological kind *river-stuff*. The correct interpretation will also depend in part on the specific details of the hypothetical scenario you're considering as actual. If it turns out that, appearances to the contrary, the drinkable stuff you call 'water' is in no way important to sustaining life, this may militate against interpreting water to be the nu-

tritional kind, *basic drink*. Of course, any interpretation of these sorts of hard case is bound to be controversial. But you get the idea. What I'm after here is a sense of the different kinds of interpretation a rational subject might come to accept when fully informed about the nature of her actual environment and of the kind of reasons which she might take to favor one interpretation over another.

As these hard cases make clear, we have no simple algorithm for determining what counts as water in every possible world considered as actual. The point becomes vivid when we look at the 2-D matrix we've constructed so far for your term 'water':

Consider the diagonal of this matrix. According to the simple model suggested by Twin Earth cases, you fill in a 2-

| ☐☐☐ | W ₁ | W ₂ | W ₃ | W ₄ | W ₅ | W ₆ | W ₇ | W ₈ | W ₉ | W... |
|--------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|---------------------------|------|
| ☐☐☐W* ₁ | H ₂ O | H ₂ O | H ₂ O | H ₂ O | H ₂ O | H ₂ O | H ₂ O | H ₂ O | H ₂ O | ... |
| ☐☐☐W* ₂ | XYZ | XYZ | XYZ | XYZ | XYZ | XYZ | XYZ | XYZ | XYZ | ... |
| ☐☐☐W* ₅ | prime matter | prime matter | prime matter | prime matter | prime matter | prime matter | prime matter | prime matter | prime matter | ... |
| ☐☐☐W* ₄ | slime | slime | slime | slime | slime | slime | slime | slime | slime | ... |
| ☐☐☐W* ₅ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | H ₂ O-or - XYZ | ... |
| ☐☐☐W* ₆ | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | (H ₂ O) (XYZ) | ... |
| ☐☐☐W* ₇ | river-stuff | river-stuff | river-stuff | river-stuff | river-stuff | river-stuff | river-stuff | river-stuff | river-stuff | ... |
| ☐☐☐W* ₈ | basic drink | basic drink | basic drink | basic drink | basic drink | basic drink | basic drink | basic drink | basic drink | ... |
| ☐☐☐W* ₉ | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | (drink) (river) | ... |
| ☐☐☐W*... | | | | | | | | | | ... |

D matrix by first relying on a simple algorithm for determining which actual samples fall into the extension of the term for any possible context of use, and then proceed to fill in the rest of each horizontal row by relying on a criterion for being the same stuff as in the samples in the actual extension. So according to the simple model, you've got a context-invariant similarity metric for determining whether something gets included in the diagonal intension: it must be the explanatory kind which satisfies whatever role you currently associate with 'water'. But just look at this matrix. What dimension of similarity unites the cells along the diagonal in this case? Some boxes along the diagonal specify chemical kinds like H_2O and XYZ, others boxes specify alien physical kinds like prime matter, others specify disjunctive kinds, biological kinds, nutritional kinds, and ecological kinds. We even have boxes that assign more than one extension to the word. In fact, there seems to be no common-sense dimension of similarity that unites all these disparate kinds of kinds: the property expressed by this diagonal intension looks utterly gruesome!⁷

Our brief examination of the intuitively correct extension of 'water' in various scenarios suggests that (i) there is no determinate sortal 'stuff' which constrains the nature of the referent, (ii) there is no strict uniqueness constraint, and (iii) there is no requirement that the actual referent must satisfy all aspects of the watery-role determined by the subject's initial attitudes and cognitive dispositions. In many cases, moreover, reaching a verdict about the extension of your

⁷Of course, that's not to say that we have *no* way of identifying the property expressed by the diagonal. It's just that we cannot directly "cotton on" to the diagonal property of this 2-D matrix for 'water'. We can only determinately identify gruesome properties indirectly—under a description, as it were. Just as we identify the property of being grue by description as *the property of being either (examined before 2001 and green) or (examined after 2001 and blue)*, we can identify the gruesome diagonal property by description as *the property expressed by the diagonal of the infinite 2-D matrix for this subject's current naïve understanding of the word 'water'*.

word in hypothetical circumstances is an extremely demanding hermeneutical exercise. These observations should lead us to reject the simple model of our ability to fill in the cells of a diagonal intension: we have no algorithm, linguistic or non-linguistic, that determines whether something falls into the extension of a word like 'water'.⁸

But then what explains our ability to fill in 2-D matrixes for such representations? How can you answer Jackson's and Chalmers's diagnostic questions about your term 'water', if you don't rely on a simple algorithm? The examples we've considered suggest that you resort to a kind of rationalizing self-interpretation. As a responsible epistemic agent reflecting on your own representational practice in the light of new and surprising information about your environment, you try to identify a referent that allows you to see the most important aspects of your practice as both reasonable and successful in that context. In other words, you try to *make the best* of your own representational practice. The philosophical literature has been dominated by distinctively third-person approaches to rationalizing interpretation, which seek to characterize the semantic content of another person's words in terms of the best rationalizing interpretation of their representational practice relative to the facts about their actual environment.⁹ But norms of ration-

⁸A number of theorists have raised these sorts of points as objections to Jackson and Chalmers. In "Conceptual Analysis, Dualism, and the Explanatory Gap", Ned Block and Robert Stalnaker target the uniqueness constraint and the watery-role constraint (pp. 14-17), and they also emphasize the role of empirical theorizing in determining the extension of a representation (pp. 23-5). In "Bad Intentions", Alex Byrne and James Pryor focus on the watery-role constraint (p. 8) and argue that Kripke's arguments from ignorance and error suffice to refute Chalmers's position. In "The Limits of Conceptual Analysis", I argue that the sortal constraint is crucial to vindicating a rigidified analysis and that there is no well-defined sortal constraint that vindicates all our intuitions about what we could learn about the referent of 'water'. All these criticisms rely on intuitions about the referents of your words in hypothetical contexts of use.

⁹W. V. O. Quine's *Word and Object*, Donald Davidson's "Mental Events", Daniel Dennett's "Intentional Systems", and David Lewis's "Radical Interpreta-

alizing interpretation can be applied to one's own representational practices as well. We often refine our own understanding of the nature of the objects, kinds, and properties we think about in the light of new and surprising information about our environment. This is, in effect, what Jackson's and Chalmers's diagnostic questions ask us to do. And the natural way to approach these questions is via norms for rationalizing interpretation. What I'm suggesting is that charity begins at home: we justify our opinions about the precise extension of our words and thoughts by trying to vindicate the most important and successful aspects of our ongoing representational practice.

Jackson and Chalmers should find this idea congenial. After all, a normal subject's initial understanding of a word like 'water' will be quite complex, involving a large range of different expectations, recognitional and inferential dispositions, and explicit opinions. The subject will also have a variety of practical and theoretical interests subserved by her ongoing categorizing practice with the term. Unlike many theorists, Jackson and Chalmers do not claim that we must sift through this complex initial understanding to find some core set of commitments that the subject should treat as settling the extension of her word no matter what the actual world turns out to be like.¹⁰ Rather, they follow the approach initially sketched by David Lewis in "How to Define Theoretical Terms": the entire complex of the subject's initial understanding of the term is taken into account in specifying its extension. Given the vast number of these tacit and ex-

tion" all characterize rationalizing interpretation from the point of view of an outside observer. Indeed, it's not clear whether their specific proposals about the nature of the process could be directly applicable from the first-person perspective. However, see Michael Root's "Davidson and Social Science" and Richard Moran's "Interpretation Theory and the First Person" for two different approaches to reconciling the first- and third-person perspectives in rationalizing interpretation.

¹⁰For this contrasting approach, see, e.g., Christopher Peacocke's *A Study of Concepts*.

plicit opinions, there will be few worlds that would vindicate all of them. When there is no perfect satisfier of the subject's naive understanding, according to Lewis, we should look for the best "near-enough" satisfier. Although Lewis doesn't offer a detailed account of what counts as a best satisfier, we may assume that he would appeal to his theory of radical interpretation to resolve such cases.¹¹ Where Jackson and Chalmers depart from Lewis is in their focus on the subject's first-person judgments about the extensions of her own representations. My proposal is that we can subsume this first-person perspective within Lewis's general framework of rationalizing interpretation: what subjects do when confronted with hard cases is engage in rationalizing *self*-interpretation.

This approach suggests a different model of a subject's ability to fill in the cells of a 2-D matrix for her representation. On this *sophisticated model*, you must take into account the totality of your initial attitudes in order to arrive at a decisive verdict about the extension of your term 'water' in a particular context of use. These verdicts must always be justified on the basis of some subset of your original attitudes and dispositions—the ones that are most important relative to that context of use. However, as we've seen, different subsets may turn out to be crucial in different empirical contexts. This means that there is no shortcut for identifying instances of the diagonal property: on the sophisticated model, each possible context of use must be considered separately, on a case by case basis.

My aim in this section has been to show why the simple model cannot explain our ability to fill in a 2-D matrix for representations like 'water'. Our best intuitions about how to answer particular diagnostic questions show that we don't rely on a simple algorithm to identify the extension of 'water' in all possible contexts of use. Instead, we must rely on holistic rationalizing self-interpretation.

¹¹See his "How to Define Theoretical Terms", "Reduction in Mind", and "Radical Interpretation".

2. A tension in 2-D semantics

Does the sophisticated model of the abilities that ground the assignment of a 2-D matrix threaten Jackson's and Chalmers's 2-D semantics? Diagonal intensions are supposed to play two key theoretical roles: they capture a priori accessible extension conditions for a token representation and they give the identity conditions for meanings and thought contents. *Prima facie*, the failure of the simple model seems to threaten both claims. In fact, however, the situation is somewhat more complex.

Consider the epistemological thesis. Jackson's and Chalmers's outspoken championship of a priori conceptual analysis and their reliance on rigidified definite descriptions in characterizing the content of terms like 'water' has led many of their critics to assume that they are committed to something like the simple model of the psychological facts that ground the ascription of 2-D matrixes. The simple model is, after all, a relatively conservative modification of traditional Fregean psychology: instead of grasping an algorithm that determines the essential nature of the referent, you grasp an algorithm that allows you to identify the referent on the basis of empirical facts about your actual environment. Given their allegiance to a broadly Fregean semantics, it's natural to interpret Jackson and Chalmers as embracing something along the lines of the simple model of our ability to fill in a 2-D matrix.

In a recent joint response to critics, however, Jackson and Chalmers have clarified their epistemological position, emphasizing how minimal the claim about a priori accessibility really is.¹² They hold that we have a priori access to the extension conditions of our representations *only* in the sense that we could in principle arrive at a correct verdict about each hypothetical context of use, considered separately,

¹²"Conceptual Analysis and Reductive Explanation" is a response to some of the criticisms raised by Block and Stalnaker in "Conceptual Analysis, Dualism, and the Explanatory Gap".

without relying on any empirical claims about our *real* environment.¹³ Jackson and Chalmers stress that they are *not* committed to there being any short definition in natural language that summarizes these a priori accessible extension conditions. Furthermore, they acknowledge that a correct specification of the extension conditions may be highly disjunctive. Indeed, nothing in their official position rules out the possibility that a correct definition for the term 'water' might require an infinitely long disjunction specifying the extension relative to each possible context of use.¹⁴ On this weakened interpretation, Jackson's and Chalmers's epistemological thesis places no constraints whatsoever on the psychological abilities that underwrite the assignment of 2-D matrixes to a subject's representations. Understood in this way, the epistemological thesis is compatible with the sophisticated model.

However, problems arise when we consider the role diagonal intensions are supposed to play in giving the identity conditions for meanings and thought contents. According to Jackson and Chalmers, two representations share the same meaning or content just in case they are associated with the same diagonal intension. In the remaining sections of the paper, I'll argue that diagonal intensions assigned on the basis of holistic rationalizing interpretation cannot play this individuating role. There are two major difficulties for the individuation thesis. First, diagonal intensions assigned in this way will be highly sensitive to idiosyncrasies in an individual's overall state of understanding. As a consequence, it will be rare that two individuals (or a single individual over time) will associate the same diagonal intension with their

¹³Since all the empirical information that might be relevant to a decision will be included in the description of the context to be considered as actual, information about one's *real* actual environment will be irrelevant to justifying a verdict about the hypothetical scenario (pp. 332-33).

¹⁴See pp. 322-23. In footnote 6, they note that Jackson is more sanguine about the possibility of finite analyses than Chalmers.

representations. Second, diagonal intensions assigned in this way will not correspond to any subjectively salient constellation of attitudes and cognitive dispositions that guides the subject's everyday use of a representation. As a consequence, subjects' everyday communication and deliberation will not be sensitive to the representational state types individuated by diagonal intensions. This instability and lack of salience, I'll argue, make diagonal intensions assigned on the basis of rationalizing interpretation ill suited to playing the role of meanings and thought contents in explaining communication and reasoning. In other words, diagonal intensions assigned on the basis of the sophisticated model cannot individuate meanings and thought contents.

Thus there is a tension between the epistemological and individuating roles that Jackson and Chalmers take diagonal intensions to play. As long as we remain at the level of the 2-D formalism, it seems plausible that diagonal intensions can fulfill both roles. But once we turn our attention to the actual psychological capacities required to play each role, it becomes clear that they are not jointly satisfiable. The psychological capacities that individuate meanings and thought contents must be relatively stable through changes in belief and they must be cognitively salient to naive subjects. But the psychological capacities that ground the subject's best judgments about the extension of representations like 'water' for all possible contexts can satisfy neither of these constraints. Close attention to the psychological interpretation of the 2-D framework will thus allow us to appreciate why two different theoretical constructs must be posited in order to fill the epistemological and individuating roles Frege assigned to senses.

3. *Public language meanings*

Can diagonal intensions assigned on the basis of holistic interpretation explain public language meanings? Frank Jackson believes they can: "in general, it is the A-intension we

know in virtue of understanding a sentence".¹⁵ A-intensions, of course, are just the intensions expressed by the diagonal of Jackson's 2-D matrixes. So Jackson's suggestion is that in order to count as competent with the public language meaning of an expression like 'water', speakers must associate a particular diagonal intension with the term. Jackson embeds this account of linguistic competence within a conventionalist approach to public language meaning along the lines sketched by David Lewis in *Convention*. Roughly, a word acquires a public language meaning when the majority of speakers converge in associating the very same information with the word, and when this convergence is mutually obvious to all speakers.¹⁶ According to Jackson, this community-wide convergence on a diagonal intension explains *what it is* for a word to have a particular public meaning. Convergence with one's community is not just a condition on an individual's competence with a word—this convergence plays a constitutive role in determining the public meaning of the word.

But is this right? Do diagonal intensions really capture the conventional linguistic meaning of expressions like 'water'? There is good reason to be skeptical. If diagonal intensions are assigned on the basis of rationalizing interpretation, it will be extremely unlikely that different members of a single linguistic community will converge on precisely the same diagonal intension.

The problem stems from the holistic character of rationalizing interpretation. In assigning an extension for a particular context, we need to take into account the subject's total understanding of the word 'water'—that is, all her inferential and recognitional dispositions, her explicit beliefs and desires, her epistemological and metaphysical assumptions, her various practical and theoretical interests in her categorizing practice, and so on. But different members of a lin-

¹⁵From *Metaphysics to Ethics*, p. 76.

¹⁶See "Reference and Description Revisited".

guistic community will inevitably associate somewhat different sets of attitudes, dispositions, and interests with a particular word. (Indeed, there would be little point in communication if they did not.) Now, in Jackson's 2-D framework, a diagonal intension is defined on every possible "centered" world considered as actual.¹⁷ The trouble is that the plentitude of different possible worlds maximizes the chances that there will be *some* centered world relative to which ideal rationalizing interpretation will assign distinct extensions to different speakers' words. And of course, all it takes for two subjects to associate distinct diagonal intensions with the word 'water' is that there be one possible context of use relative to which their words are assigned distinct extensions. So if it's true that diagonal intensions are assigned via holistic rationalizing interpretation, the odds are heavily stacked against convergence on the same diagonal intension.

An example might help make this point clearer. Let's imagine a Chicago commodities broker and his naive new-age neighbor. The broker is passionately interested in the droughts affecting harvests and transportation, and he spends hours discussing rain and profits with cronies over beers late into the night. His neighbor, in contrast, devotes her time to chauffeuring her kids to swimming and soccer, doing household chores, practicing yoga, and reading up on the simplicity movement. Whereas the broker survives on coffee and beer, the new-ager is fanatical about bottled water. Despite these differences, however, the two neighbors express many of the same opinions about the stuff they call 'water', and they are both well within the norms for minimal linguistic competence with the term.

Do they associate precisely the same diagonal intension with the word 'water'? I doubt it. Because of their divergent interests, it's plausible that there will be some actual world candidates for which rationalizing interpretation

¹⁷See "Conceptual Analysis and Reductive Explanation" for his most recent formulation of the position.

would assign distinct extensions to their words. Consider a context which falsifies the neighbors' shared assumption that there is a single scientific kind underlying the various samples they standardly count as 'water': instead of H₂O, the lakes, rivers and aquifers are filled with a variety different chemical compounds—X₁, X₂,..., X₁₀. In contexts like this, it's plausible to assign a functional kind as the referent of the subjects' words. But do their words refer to precisely the *same* functional kind in this context? In order to address this question, we need to consider which functional roles would be most important to each subject's own concerns and practices with the word. In the present case, the task is facilitated by the fact that our two subjects have relatively well-defined interests. The commodity broker's dominant interest is in the role of rain in irrigating crops and filling rivers. It seems reasonable, therefore, to assign an ecological kind, 'rain-stuff' (i.e. whatever falls as rain and irrigates crops) as the semantic value of his word relative to this context of use. Since the new-ager's main interests in categorizing stuff as 'water' revolve around its nutritional and symbolic roles, we should assign her word a functional kind which reflects these interests—something like 'basic drink'. But there will be actual world candidates relative to which the extension of these two functional kinds, rain-stuff and basic-drink, will diverge. For instance, suppose that the actual world contains one compound, X₁, which fulfills the ecological role, but not the nutritional role. So samples of X₁ belong to the ecological kind river-stuff but they don't belong to the nutritional kind basic-drink. Relative to this world, the extension of the word 'water' as used by the two neighbors will be different: the extension of the broker's term will include samples of X₁, whereas the new-ager's term will exclude them. It follows that the neighbors associate different diagonal intensions with the word.

In the interests of cutting short hermeneutical debates, the example is stylized. However, the moral is a general

one. If diagonal intensions are part of the public meaning of 'water', then at least one of the two neighbors must fail to grasp the word's meaning. But this is implausible: intuitively, both should count as competent with the public meaning of the word. More importantly, the example gives us good reason to doubt that diagonal intensions could form any part of the public meaning of the term. According to the conventionalist account of public meanings, the majority of English speakers must associate the same information with the term and this fact must be mutually obvious to all speakers. But if these two neighbors fail to converge on a single diagonal intension, it's hard to believe that English speakers will in general succeed in so converging. And it's even less plausible that this fact would be mutually obvious to all concerned. However, if there is no widespread and mutually obvious convergence on a single diagonal intension for 'water', then by Jackson's own lights the diagonal does not belong to the public meaning of the expression. Although diagonal intensions assigned on the basis of holistic rationalizing interpretation of individual speakers' total state of understanding may capture a priori accessible extension conditions which individual speakers associate with their words, they cannot be what individuates public language meanings.¹⁸

I'll consider two responses to this line of argument. The first response claims that linguistic deference guarantees that most speakers will converge on a single diagonal intension.

¹⁸In "Conceptual Analysis and Reductive Explanation", Jackson and Chalmers acknowledge that different speakers might come to divergent verdicts about the extension of names like 'Neptune', and they raise the possibility that this might hold true of natural kind terms like 'water' (pp. 327-28). That paper, however, is concerned with their epistemological thesis. They simply note that divergent diagonal intensions do not threaten the claim that each individual speaker has "in principle" a priori access to facts about the extension of her own term in each possible context of use. They do not consider what consequences this might have for the use of 2-D semantics to individuate public meanings. Unlike Jackson, Chalmers is skeptical that diagonal intensions can always capture public meanings (see "On Sense and Intension", section 8).

The second response admits that speakers will not converge, but claims that it's the similarity among individual speakers' diagonal intensions that determines the public language meaning of words.¹⁹

Let's consider linguistic deference phenomena. All speakers have a strong interest in using words in the same way as others in their community. Moreover, normal speakers naturally assume that they are co-referring with other members of their linguistic community. And they are usually willing to some degree to be persuaded by others that their own use of an expression is incorrect on linguistic or epistemological grounds. In my account of the two neighbors, however, I did not focus on these social interests and dispositions. One might suspect that when these facts are taken into account, rationalizing interpretation of the two neighbors would assign the very same diagonal intension to both speakers' use of the word 'water'. Indeed, it may seem that the prevalence of these sorts of background interests and dispositions will be enough to guarantee that the majority of English speakers would converge on the very same diagonal intension for the word.

In fact, however, the basic problem remains unchanged when we factor in linguistic deference phenomena. When we include social interests and dispositions in the speaker's initial understanding of a word, we need to consider the relevant social facts as part of the different possible contexts of use. So in comparing the diagonal intensions the two neighbors associate with 'water', we need to consider how each speaker should be interpreted relative to all possible social environments. But different speakers will have slightly different interests and dispositions with regard to their linguistic community. For instance, the broker may be inclined to trust different sorts of people than the new-ager;

¹⁹In "Reference and Description Revisited", Jackson seems to invoke both lines of response. For the linguistic deference response, see pp. 208-9, for the similarity response, see his discussion of names on pp. 214-15.

and he may have more of an interest in coordinating his use of the word with others' than she does. These sorts of differences will allow us to identify social environments relative to which ideal rationalizing interpretation would assign distinct extensions for the two neighbors' use of the word 'water'. So even when we take social deference phenomena into account, it turns out that the two neighbors associate distinct diagonal intensions with the word.

Let's turn now to the second response to the argument. Even if there is no single diagonal intension for 'water' that most English speakers will converge on, it seems plausible that different speakers' diagonal intensions will overlap for a great many contexts. After all, most people seem to agree that if the actual world is as chemistry tells us, then water is H_2O , and if the actual world is like Twin Earth, then water is XYZ. So perhaps Jackson's claim should simply be weakened: the public linguistic meaning of 'water' is determined by the *similarity* of the diagonal intensions speakers associate with that word. One way of formalizing this proposal is to say that the public meaning of the word 'water' is given by a *partial* diagonal intension, which assigns extensions only to a limited subclass of all possible contexts of use. The 2-D theorist would then need to specify the relevant subclass of contexts in such a way as to capture an overlap in how most English speakers understand the word 'water'. Since partial diagonal intensions would be much easier to share than full diagonal intensions, this approach would avoid the problems with interpersonal convergence that plague Jackson's original proposal.

It's important to notice that the partial-diagonal proposal constitutes a renunciation of the original ambitions of Jackson's 2-D semantics. The idea we've been examining is that there is a single theoretical construct—a diagonal intension—that serves both to determine a priori extension conditions and to individuate public-language meanings. According to the current proposal, however, nothing plays

both of these theoretical roles. Complete diagonal intensions determine the a priori extension conditions of a subject's words, but they don't individuate public meanings. Partial diagonal intensions individuate public meanings, but they don't fully determine the a priori extension conditions of individual speakers' words.²⁰ Thus, the partial-diagonal account cannot vindicate a Fregean account of public meanings, since nothing plays both the epistemological and the individuating roles that are distinctive of Fregean senses.

Even if it fails to vindicate Jackson's Fregean ambitions, however, the 2-D framework might still play a crucial role in individuating public meanings. It's worthwhile taking a closer look at the partial-diagonal proposal to see whether this promise is borne out.²¹ On the sophisticated model, diagonal intensions provide fine-grained information about individual speakers' understanding of a term. Moreover, they present this information in a very simple format, as a function from possible contexts of use to extensions. By limiting the range of possible contexts over which this function is defined, we can specify highly abstract commonalities in speakers' understanding of a word—commonalities that may be beyond the reach of the coarser conceptual framework of commonsense psychology. If partial diagonal intensions really do allow us to capture precisely what's required for competence with the public meaning of a word, this would constitute a significant vindication of the 2-D framework. The technical tools of 2-D semantics could then play an important role in clarifying the relevant kind and degree of similarity in different speakers' understanding required in order for them to share a public meaning.

²⁰On the partial-diagonal account of public meaning, two speakers who associate the very same (non-indexical) meaning with the word 'water' may nonetheless use the term to refer to distinct kinds of stuff. For instance, if the actual world contains no unified explanatory kind underlying watery phenomena, the two neighbors may refer to distinct functional kinds with the term 'water'—river-stuff and basic-drink—even though they share the same public meaning.

²¹Thanks to two anonymous referees for pressing me to clarify this point.

In order to make good on this promise, the proponent of partial diagonals must say how exactly the class of relevant alternative contexts of use must be restricted so that partial diagonals will capture public meanings. Obviously, public meanings cannot be identified with just *any* overlap in the diagonal intensions most speakers associate with the term—for some kinds of overlap have nothing to do with conventional meaning. Although the terms 'renate' and 'cordate' do not share the same public meaning, they share a partial intension: in all scenarios that are biologically similar to the actual world, creatures with hearts always have kidneys, and vice versa. In general, any contingently co-extensive terms will have overlapping diagonal intensions for a significant range of relevant alternative contexts. In order to avoid this sort of trivial counterexample to the partial-diagonal account of public meanings, the class of relevant alternative contexts on which partial diagonal intensions are defined must be restricted so as to reflect speakers' own intuitions about when words have the same or different meanings. After all, the sameness or difference of public meanings must be publicly accessible to ordinary speakers. Moreover, a version of this subjective accessibility constraint is built into the conventionalist account of meaning: in order for there to be a tacit convention associating a certain pattern of understanding with a word, it must be *mutually obvious* to all speakers that most of them understand a term in that way.

Thus, if partial diagonals are to capture tacit linguistic conventions, it must be mutually obvious to all speakers that most of them converge in the kind of understanding specified by partial diagonal intensions. It would, however, be absurd to insist that ordinary English speakers recognize that they share the same partial diagonal intensions *under that description*. The machinery of 2-D semantics is a theorist's construct, and a very recent one at that. So what sorts of mutually obvious facts about speakers' ordinary ways of

understanding a word could ground the assignment of a partial diagonal intension?

It might seem that there is a very simple response. Perhaps it's mutually obvious to all English speakers that most of them would converge in their (independent) verdicts for a specific range of hypothetical conditionals of the form, 'if the actual world turns out to be thus-and-so, then water is...'. These conditional verdicts can then be used by a 2-D theorist to define a partial diagonal intension that captures this mutually obvious range of convergence. Thus, there's no need to assume that ordinary speakers know what 'partial diagonal intension' means. However, this simple response will not do, since it still builds in too much conceptual sophistication into ordinary speakers' understanding of 'water'. The hypothetical scenarios that define a 2-D matrix are complete microphysical descriptions that characterize the state of the world down to the very last gluon. Ordinary English speakers like you and me are incapable of understanding such complicated descriptions, much less of using them to enumerate all the microphysical complexes that fall into the extension of the term 'water'. Hypothetical scenarios and extension assignments are theoretical notions that form no part of speakers' ordinary understanding of what a word represents. Moreover, as we've seen, arriving at a verdict about the extension of an individual speaker's word for a given scenario can be an extremely delicate task involving ideal rationalizing interpretation of the speakers' total state of understanding. Given our imperfect knowledge of others' initial attitudes and dispositions, this interpretive task will be much more difficult in the third-person case than in the first-person case. Thus the suggestion that ordinary speakers are implicitly aware of a convergence in extension assignments for a particular range of hypothetical scenarios is psychologically unrealistic.

If partial diagonal intensions are to capture mutually obvious standards for correct linguistic usage, they must reflect

ordinary commonsense ways of understanding words. So instead of positing mutually obvious patterns in hypothetical extension assignments, the proponent of partial diagonals should instead look for mutually obvious patterns in the attitudes and cognitive dispositions that constitute a normal speaker's everyday understanding of what her word represents. We can think of this sort of pattern as something like an implicit commonsense stereotype of the referent. I take it that a proponent of the partial-diagonal account can meet the mutual obviousness constraint only if partial diagonal intensions capture something like conventional stereotypes that ordinary speakers treat as necessary and sufficient for competence with the meaning of words.

There are two sorts of objection one might raise at this point. One might reject the suggestion that public meanings are individuated by something like commonsense stereotypes. After all, it's not clear that names and natural kind terms are always associated with a conventional stereotype of the referent. And even when they are, it's doubtful whether ordinary speakers treat acceptance of the stereotype as necessary and sufficient for competence with the meaning of the word. However, these are general objections to Lewisian conventionalism that I want to set aside for present purposes. My concern here is specifically with the use of Jackson's 2-D framework to capture conventional linguistic meanings. This technical framework, I'll argue, lacks the expressive resources to isolate the kind of stereotypical understanding which could be mutually obvious to ordinary speakers.

If they are to figure in tacit linguistic conventions, commonsense stereotypes cannot be too complex or difficult to discern. Stereotypes can involve only a limited constellation of cognitively shallow and intersubjectively salient attitudes and dispositions. But diagonal intensions are designed to capture a very different kind of information about a speaker's understanding: they encode abstruse, cognitively

inaccessible facts about the ideal upshot of holistic interpretation of the speaker's word for every context of use. Commonsense stereotypes will, of course, be part of the *input* into the process of ideal rationalizing interpretation—but so will all of the speaker's other attitudes and cognitive dispositions. In many contexts, the speaker's cognitively shallow stereotypes will be overridden by her other cognitive commitments. As a result, information about stereotypes will be lost in the process of ideal holistic reflection.

An example might help make this point clearer. Like many first year philosophy students, Max implicitly assumes that 'free will' applies to an action only if the action was not causally predetermined by any physical processes. We may assume that this libertarianism forms part of Max's implicit stereotype of free will. However, ideal a priori reflection (independently of any assumptions about the context of use) might well lead Max to reject libertarianism in favor of compatibilism. In that case, the stereotype Max originally associated with 'free will' would not figure in a justification of his ultimate verdicts about the extension of the term in any possible context of use. But then the resulting diagonal intension would convey no information about the libertarian aspect of Max's original stereotype—for someone who failed to share Max's initial libertarian intuitions might nonetheless share the very same diagonal intension for 'free will'. The moral is that diagonal intensions are the wrong technical tool to capture information about stereotypes: information about initial cognitively salient attitudes and dispositions may simply be "washed out" in the idealized holistic reflection that determines the assignment of diagonal intensions.

However, moving from full diagonal intensions to partial diagonal intensions will only exacerbate the problem we've been considering. As we've seen, the full diagonal intensions that individual speakers associate with a word will lose information about the original stereotype. Partial diagonal intensions, however, convey even *less* information about the

speaker's original cognitively shallow dispositions than do full diagonal intensions. A fortiori, partial diagonal intensions are ill suited to isolate speakers' shared commonsense stereotypes. We should therefore reject the suggestion that the theoretical apparatus of Jackson's 2-D framework plays an important role in clarifying what it takes to share a public meaning. On the sophisticated model, the 2-D framework provides the wrong technical tools for capturing mutually obvious features of naive speakers' shared understanding.

In this section, I've argued that the identity conditions of conventional linguistic meanings cannot be explained in terms of sophisticated diagonal intensions. Conventional meanings must capture a mutually obvious convergence in the way most speakers understand a word. The example of the two neighbors shows that it's unreasonable to expect different members of a linguistic community to converge independently on the very same diagonal intension. It might seem that this problem could be mitigated by taking speakers' "social deference" dispositions into account. However, an extension of the example shows that this move does not in fact reduce the divergence in diagonal intensions among members of a linguistic community. Lastly, I considered whether meanings might be individuated by partial diagonal intensions. This would be to abandon Jackson's original Fregean program. I argued, moreover, that partial diagonal intensions do not provide the expressive resources to isolate what ordinary speakers find mutually obvious. I conclude that two-dimensional semantic values assigned on the basis of holistic rationalizing interpretation are ill suited to individuating public language meanings.

4. *Units of cognitive significance*

In "The Components of Content", Chalmers argues that diagonal intensions provide a good theoretical characterization of the type of fine-grained representational state invoked by commonsense psychology in assessing a subject's

rationality or explaining her reasoning. In particular, he claims that diagonal intensions can account for our commonsense intuitions about six well-known philosophical puzzle cases. Here's a brief recap:²²

1. Putnam's Twin Earth: What do we have in common with our Twin Earth counterparts?
2. Frege's non-trivial identity claims: Why is 'Hesperus is Hesperus' trivially true whereas 'Hesperus is Phosphorus' is potentially informative?
3. Kripke's Pierre: How can Pierre fail to be irrational when he believes, of London, that it is pretty and that it is not pretty?
4. Perry's essential indexicals: Why does accepting 'I am making a mess' have an immediate effect on Perry's actions, whereas 'Perry is making a mess' need not?
5. Modes of presentation: What is it about Lois's representational states that makes 'Lois believes Superman can fly' true when 'Lois believes Clark Kent can fly' is false?
6. Contingent a priori knowledge: How can I know a priori that a certain stick in Paris is one meter long?

²²For the classical statement of the puzzles see: Hilary Putnam's "The Meaning of Meaning", Gotlob Frege's "On Sense and Reference", Saul Kripke's "A Puzzle about Belief", John Perry's "The Problem of the Essential Indexical", Stephen Schiffer's "The Mode of Presentation Problem", and Saul Kripke's *Naming and Necessity*.

As these cases illustrate, our understanding of a person's representational states is not exhausted by our grasp of his ordinary wide content. Commonsense psychology invokes finer-grained representational states in order to underwrite assessments of the subject's rationality (as in puzzles 2 and 3) and explanations of his inferences and behavior (as in puzzles 1 and 4). For instance, whether Kripke's Pierre is rationally blameworthy in accepting beliefs with contradictory wide contents depends on whether he redeploys the very same representational state in his two beliefs. Similarly, whether Perry will become embarrassed and start cleaning up the mess he's making (causally) depends on which representational states are involved in his belief with the wide content *Perry is making a mess*. Let's call the representational states that play the core normative and explanatory roles in commonsense psychology *the units of cognitive significance*. Chalmers's suggestion is that diagonal intensions give the identity conditions for these commonsense units of cognitive significance: two token representations (be they mental or linguistic) express the very same unit of cognitive significance just in case those tokens have the very same diagonal intension.

If Chalmers is right, diagonal intensions fulfill both the epistemic and semantic roles of Fregean senses: they give a priori extension conditions for the subject's token representations, and they provide the identity conditions for fine-grained representational states invoked in commonsense psychology.²³ I'll argue, however, that representational states whose identity is determined by diagonal intensions cannot fulfill the core normative and the explanatory roles in commonsense psychology that Chalmers highlights.

²³Chalmers argues that diagonal intensions are a plausible successor to Fregean senses in "On Sense and Intension". The two theoretical roles I focus on correspond roughly to Chalmers's second and fourth conditions on Fregean senses (section 2). I'll concentrate here on Chalmers's arguments in "The Components of Content", since they are more detailed and largely subsume the considerations he takes to be relevant to individuating Fregean senses.

Let's first consider the normative role. Chalmers claims that diagonal intensions can characterize what he calls "rational relations" between thoughts. Whether particular token thoughts, or combinations of thoughts, are rationally required or forbidden depends on whether the same units of cognitive significance are redeployed in those thoughts. Chalmers cites three examples: the units of cognitive significance determine whether an identity claim is trivial or potentially informative (puzzle 2); they determine whether beliefs with contradictory wide contents are rationally blameworthy (puzzle 3); and they determine whether a subject can know that a contingent claim is true or false wholly a priori (puzzle 6). I'll concentrate here on Chalmers's first two examples of rational relations—non-trivial identity and blameless contradictions. These are the sorts of examples that originally led Frege to postulate a second level of semantic content in "On Sense and Reference", and such examples have played a central role in the philosophical debate over the identity conditions of fine-grained representational states since that time.²⁴

What's distinctive of rational relations like triviality and non-contradiction is that they involve extremely minimal standards of rationality. Consider triviality. We simply cannot make sense of the idea that someone might literally believe that Hesperus is not identical to Hesperus. According to commonsense psychology, it's obvious from the thinker's own point of view that a thought of the form $[A \neq A]$ is about a single object failing to be self-identical. You'd have to be crazy or malfunctioning in order to even consider accepting a belief of that form. Empirical ignorance is no excuse, nor is a lack of time for careful reflection—it should simply be obvious to you that you're thinking about a single

²⁴I won't take a stand here on whether diagonal intensions can explain a priori knowledge of contingent propositions—though I have my doubts. Unlike Chalmers, however, I am not committed to the idea that all six of his puzzle cases must be explained by the very same representational states.

object. This is emphatically not the case for non-trivial identity claims: you may be ill advised or careless in thinking that Hesperus is not identical to Phosphorus, but your thought isn't crazy. This, I suggest, is the fundamental difference between trivial and non-trivial identity claims: a thinker who doubts a trivial identity claim violates minimal standards of rational intelligibility.

Our reaction to Kripke's Pierre can be explained along similar lines. During the course of a multi-lingual party, Pierre tells us, with a great air of sincerity, "Londres est jolie" and then immediately adds "London is not pretty". How should we interpret him? Surely Pierre cannot be using the words 'Londres' and 'London' to express the very same unit of cognitive significance. If that was what was going on, then Pierre would have committed himself to a blatant contradiction! But Pierre doesn't seem to have lost his mind. Nor does he seem to have forgotten his earlier approval of Londres when he is running down London. So, we conclude, he must associate distinct units of cognitive significance with his words 'Londres' and 'London'. That is, we assume his contradictory thoughts have the logical form $[Ps \ \& \ \sim Pn]$, rather than $[Ps \ \& \ \sim Ps]$. Again, what's driving our interpretation is the conviction that blatant contradictions violate minimal standards of rational intelligibility.

These commonsense judgments presuppose that the sameness or difference of the units of cognitive significance associated with mental or linguistic representations is obvious to the subject himself. That is, there must be something about the way the thinker grasps two token representations that makes it obvious and incontrovertible from his perspective that those representations are about the very same subject matter. If it took any cognitive effort to figure out that a thought of the form $[Hesperus \neq Hesperus]$ is incoherent, then there would be nothing especially problematic about someone who accepted the claim: we'd simply assume that he hadn't done all his logical homework. Similarly, it's

precisely because we think the sameness or difference of the units of cognitive significance is subjectively obvious that we take Pierre's contradictory claims to be evidence that 'Londres' and 'London' express distinct units of cognitive significance. Without this assumption, there would be nothing wrong with ascribing to Pierre thoughts of the form $[Ps \ \& \ \sim Ps]$.

Similar epistemic assumptions are implicit in commonsense rationalizing explanation of inferences. Here's an example of the kind of explanation Chalmers has in mind:²⁵

Suppose I think that Superman is across the road, and I want to have Superman's autograph: then other things being equal, I will cross the road. If you have thoughts with similar epistemic content [i.e. diagonal intension] to mine, then you will do the same. If your thoughts share only subjunctive content [i.e. horizontal intensions] with mine, while having different epistemic content [i.e. diagonal intension]—perhaps you think that Clark Kent is across the road, but want Superman's autograph—then your corresponding behavior may be quite different.

Why does only one of these subjects put his belief and desire together in a practical syllogism that issues in a rationally motivated intention to cross the road? It's clear that ordinary wide content cannot fully capture the explanatorily relevant features in this case, since both subjects have a belief and a desire with the same wide contents, and yet only one of them forms a corresponding intention. Chalmers's suggestion is that diagonal intensions can isolate the causally relevant factors in this case—or at least they do a better job than ordinary wide content.²⁶

²⁵"The Components of Content", p. 620. Chalmers is assuming that Superman is actual and is identical to Clark Kent.

²⁶Chalmers is careful not to make too strong a claim about the causal relevance of diagonal intensions. He holds that the contribution of ordinary wide

Using the very same name, 'Superman', to characterize the content of the first subject's thoughts naturally leads us to conclude that it's obvious from the subject's own point of view that his belief and desire are about the very same subject matter. This immediate appearance of sameness of subject matter is what explains why the subject is inclined to put those thoughts together in a practical syllogism (assuming, of course, that he's minimally rational). Similarly, the *lack* of this immediate appearance explains why the second subject is not so inclined. When Chalmers uses two different names, 'Superman' and 'Clark Kent', to characterize the contents of the second subject's belief and desire, we naturally assume that it's *not* obvious from the subject's own point of view that those thoughts are about the very same subject matter. Otherwise, why would a non-crazy subject fail to draw such an obvious conclusion? Whether a minimally rational subject is disposed to put two thoughts together in reasoning, I'm suggesting, depends on whether it is obvious and incontrovertible from the subject's own point of view that they concern the very same subject matter.²⁷

These observations about the normative and explanatory presuppositions behind the puzzle cases suggest a criterion of adequacy for Chalmers's account. If they are to capture

content to causal explanations is "screened off" by that of diagonal intensions. That is, if we were to hold the diagonal intension of the subject's thoughts fixed, but varied their horizontal intensions, the subject would think and act precisely the same way. But if we varied the diagonal intensions, keeping their horizontal intensions fixed, the inference or action would not be preserved. ("The Components of Content", pp. 619-20.)

²⁷If it wasn't immediately apparent to the subject that the belief and desire purport to represent the very same man, then there would have to be some *further* explanation of why he put those thoughts together in practical reasoning. Why did the subject think *this* man was *that* man? Often there will be a genuine reason we can cite in answering such questions. But the regress of reasons cannot go on forever or we'll face an analog of Lewis Carroll's problem of Achilles and the Tortoise (see Carroll's "What the Tortoise said to Achilles"). At some point, the subject's reasoning must be explained by the brute fact that certain thoughts strike him as obviously about the very same subject matter.

the units of cognitive significance in play in commonsense psychology, diagonal intensions must group token representations into types in such a way that it's immediately obvious and incontrovertible from the perspective of the thinker himself that all such token representations are about the very same subject matter. Sameness of diagonal intension, in other words, must capture subjectively apparent sameness of subject matter.

Diagonal intensions assigned on the basis of holistic rationalizing self-interpretation cannot satisfy this criterion of adequacy. On the sophisticated model, diagonal intensions record the upshot of ideal rational reflection relative to each possible context of use considered independently. Whether two representations share the same diagonal intension will thus depend on highly abstract and cognitively inaccessible features of the subject's total cognitive state with respect to each representation. As we saw in the discussion of public meanings, representational states individuated by these diagonal intensions are ill suited to capturing subjectively salient aspects of ordinary subjects' understanding of a representation.

Rational inquiry provides a particularly vivid illustration of this point. Consider what happens when a subject discovers the true nature of water through empirical inquiry. Learning about the chemical structure of water will tend to instill new theoretical beliefs, dispositions to take new kinds of evidence into account, commitments to new taxonomic systems, new explanatory ambitions, new ways of accommodating other speakers' beliefs, and so on. As a consequence, we should expect the diagonal intensions associated with a subject's token 'water' representations to change significantly over the course of the inquiry. But when you engage in rational inquiry, it strikes you as obvious and incontrovertible that the subject matter whose nature you're trying to discover remains stable throughout the process of inquiry. Indeed that's the whole point of rational inquiry:

you're trying to discover facts about the very stuff you were curious about earlier. You don't take yourself to be changing the subject every time you learn something important about water. Diagonal intensions assigned on the basis of holistic rationalizing interpretation, however, won't capture this subjectively obvious sameness of subject matter.²⁸

Diagonal intensions do capture a certain kind of subjective access to incontrovertible sameness of subject matter. If two token representations share the same diagonal intension, then the subject can "in principle" know a priori that they must represent the same subject matter. That is to say, if the subject were to compare the upshot of ideal rationalizing self-interpretation of each representation relative to each of the plenitude of hypothetical contexts of use, she could verify that both representations must have the same extension without relying on any empirical information about her real environment. The trouble is that this ideal a priori accessibility is a far cry from the immediate obviousness which is crucial to individuating the units of cognitive significance in commonsense psychology. As the case of empirical inquiry illustrates, it may strike the subject as obvious that two token representations are about the same subject matter even when they fail to share the same diagonal intension. And even when two tokens do share the same diagonal intension, there's no reason to suppose that it must be obvious to the subject herself that they represent the same thing. In view of this radical mismatch in their epistemic properties, it's simply not credible that diagonal intensions capture the identity conditions of the units of cognitive significance we appeal to in assessing subjects' minimal rationality or in explaining their reasoning.

²⁸Chalmers himself does not cite diachronic examples. However, any elucidation of our commonsense ways of individuating units of cognitive significance should generalize to diachronic reasoning. Moreover, we should expect similar problems to arise whenever the subject engages in genuinely informative empirical inferences at a particular time.

I'll close this section by considering how Chalmers might respond to this objection. Chalmers is aware that representational states individuated by diagonal intensions will not correspond *perfectly* to the commonsense units of cognitive significance which Frege was interested in. In particular, he's prepared to concede that mathematical and logical reasoning cannot be explained in terms of diagonal intensions.²⁹

This understanding of cognitive significance [i.e. in terms of diagonal intensions] is not quite Frege's. On Frege's account, a priori knowledge can be cognitively significant: the knowledge that $59+46$ is 115 is cognitively significant, for example, because this knowledge requires some cognitive work. It is very hard to articulate this notion precisely, however, and it is not clear that there is a useful precise notion nearby. In any case, the definition in terms of a priority is at least not too far from Frege's central notion, and it handles most of Frege's central cases, which involve a posteriori knowledge. So it is this understanding that I will use.

In a priori domains of inquiry such as mathematics, Chalmers concedes that diagonal intensions cannot capture commonsense units of cognitive significance: after all, ' $59+46 = 115$ ' will have the same diagonal intension as ' $115 = 115$ ' (they're both true in all possible contexts of use), but clearly you are in very different representational states when you entertain these two claims. Nonetheless, Chalmers maintains that a priori domains of inquiry constitute a relatively minor exception to his general thesis that diagonal intensions individuate commonsense units of cognitive significance.

Our discussion suggests that Chalmers has seriously underestimated the difficulties facing his account of the identity conditions of fine-grained representational states. Even when we confine our attention to thoughts about empirical

²⁹"On Sense and Intension", p. 15.

topics like astronomy or geography, diagonal intensions fail to individuate the fine-grained representational states invoked in our commonsense understanding of Chalmers's central puzzle cases. In order to vindicate our normative and explanatory practices, I've argued, the units of cognitive significance must be individuated in terms of what's immediately obvious to a minimally rational subject. What distinguishes potentially informative empirical identities like 'Hesperus = Phosphorus' from trivial ones like 'Hesperus = Hesperus' is whether the identity claim strikes the subject as obviously and incontrovertibly about the very same subject matter. This is also what distinguishes potentially informative mathematical identities like ' $59+46 = 115$ ' from trivial ones like ' $115 = 115$ '. Thus, the very same principle of individuation seems to be in play for both a priori and a posteriori domains. Insofar as there is a problem for Chalmers's account in the mathematical domain, the problem arises in empirical domains as well.

But even if this problem is quite general, Chalmers may be unimpressed. He suggests that common sense individuates the units of cognitive significance in mathematical claims in terms of how much "cognitive work" would be required in order to realize that the claim is true. Since cognitive work comes in various kinds and degrees, there may be no non-arbitrary way of specifying precisely which kind of cognitive work is relevant to individuating fine-grained representational states. So Chalmers might conclude that commonsense psychology provides no univocal, theoretically tractable standard for individuating representational states. If this is right, Chalmers needn't be too bothered that his systematic theoretical account is somewhat revisionary.

However, this line of thought provides no support for Chalmers's revisionary account of the units of cognitive significance. In the first place, Chalmers's elucidation of commonsense psychology is inaccurate. Our examination of the puzzle cases suggests that commonsense psychology indi-

viduates representational states in terms of what's immediately obvious and incontrovertible from the subject's point of view. This is an absolute standard—immediate obviousness doesn't come in different kinds or degrees. If this account of the commitments of commonsense psychology is correct, then the case for Chalmers's revisionary theory of the units of cognitive significance simply lapses: commonsense doesn't individuate representational states in terms of a vague standard of kinds and degrees of cognitive work.³⁰

Moreover, even if common sense did individuate the units of cognitive significance in terms of vague degrees of cognitive work, this fact wouldn't lend any credence to Chalmers's radical revisionism. Replacing the absolute standard of immediate obviousness with a graded notion of cognitive work would generate a proposal along the following lines: commonsense psychology individuates representational states in terms of whether subjects can discern *without too much difficulty* whether or not two thoughts are about the very same subject matter. Chalmers's diagonal intensions, in contrast, individuate representational states in terms of what the subject can "in principle" know a priori. As we've seen, something could be a priori knowable in Chalmers's sense even if it would require infinitely many exercises of ideal rational deliberation to actually know it. In effect, then, Chalmers's account idealizes away from *any* substantive cognitive work constraint—for it could take an infinite amount of cognitive work to achieve the relevant

³⁰The metaphor of mental file-folders can help elucidate the role played by the units of cognitive significance in commonsense psychology: mental files bind together evolving bodies of information (i.e. dispositions and attitudes) in such a way that they immediately strike the thinker as pertaining to a single subject matter. Thus, the units of cognitive significance capture the subject's most basic way of taking token representational states to be about the very same subject matter. We can think of these units as determining a *minimal* degree of cognitive work: in entertaining two tokens of the same unit of cognitive significance, a minimally rational subject eo ipso takes them to be about a single subject matter.

kind of a priori knowledge. It's hard to see how an account *this* revisionary could claim to be an elucidation of our commonsense notion. As we've seen, it's crucial to our commonsense normative and explanatory practices—and to the central puzzle cases Chalmers wishes to explain—that ordinary subjects be sensitive to the sameness or difference of the units of cognitive significance. In ignoring this commonsense accessibility constraint, Chalmers's account simply changes the subject under consideration.

I conclude that Chalmers's diagonal intensions cannot give the identity conditions of the units of cognitive significance invoked in commonsense psychology. Chalmers claim was that diagonal intensions could underwrite ordinary assessments of rationality and rationalizing explanations. In this section, I've argued that the epistemological properties of diagonal intensions make them ill suited to playing these roles. Immediate and incontrovertible access is one thing; "in principle" accessibility on the basis of infinite and ideally rational a priori reflection is another. The sameness or difference of the units of cognitive significance is accessible in the former way, while the sameness or difference of representational states individuated by diagonal intensions need only be accessible in the latter way.

5. Conclusion

Jackson and Chalmers think their 2-D semantic framework can vindicate a plausible successor to Frege's notion of sense. Like Fregean senses, diagonal intensions are meant to play two key theoretical roles: (i) they provide a priori accessible extension conditions for token representations and (ii) they provide identity conditions for the types of fine-grained representational states that commonsense psychology invokes in characterizing the rational evolution of thought and talk. Prima facie, diagonal intensions appear well suited to playing these roles. By recording the subject's own ideally rational commitments as to what falls into the extension of a

representation relative to all possible contexts of use, diagonal intensions capture a species of extension condition for the subject's representation. At the same time, diagonal intensions capture something of the subject's own perspective on the extension of her representations. It's tempting, therefore, to conclude that diagonal intensions can play both the epistemological and individuating roles of Fregean senses.

However, this temptation evaporates when we pay closer attention to the actual phenomena to be explained. Jackson and Chalmers seek to accommodate semantic externalism within their broadly Fregean framework. In order to do so, they claim that the extension conditions of a subject's words and thoughts are determined by ideal rational reflection on full empirical information about a context of use. However, I have argued that the identity conditions of fine-grained representational states such as meanings and thought contents depend on what's immediately obvious to a minimally rational subject. Ideal rationality is one thing; minimal rationality is another. Both kinds of rationality are important to commonsense psychology: the rational evolution of thought and talk cannot be understood without appeal to minimal standards of rationality, whereas evaluating the objective success of this thought and talk—its truth—requires a different, highly idealized standard of correctness. The moral of our discussion is that no single theoretical entity can simultaneously capture both of these standards. Even if we grant Jackson and Chalmers that their 2-D framework can accommodate the epistemological insights of semantic externalism, the resulting diagonal intensions do not succeed in bridging the gap between ideal and minimal rationality that is at the heart of commonsense psychology.³¹

³¹Material from this paper was presented at the 2001 Australasian Association of Philosophy, the Department of Philosophy at Macquarie University, and the 2002 Western Division meeting of the American Philosophy Association. I'm grateful to the audiences on those occasions and especially to Takashi Yagi-

References

- Block, N., and R. Stalnaker. (1999). "Conceptual Analysis, Dualism, and the Explanatory Gap." *Philosophical Review* 108: 1-46.
- Byrne, A., and J. Pryor. (2001). "Bad Intensions." Ms. presented at the Barcelona Workshop on the Theory of Reference II. (<http://mit.edu/abyrne/www/BadIntensions.pdf>).
- Carroll, L. (1895). "What the Tortoise Said to Achilles." *Mind* 4: 278-80.
- Chalmers, D. (2002). "On Sense and Intension." Ms. (<http://www.u.arizona.edu/~chalmers/papers/intension.html>).
- (2002). "The Components of Content." *Philosophy of Mind: Classical and Contemporary Readings*. D. Chalmers. Oxford, Oxford University Press.
- (forthcoming). "The Foundations of Two Dimensional Semantics." *Philosophical Studies*.
- Chalmers, D., and F. Jackson (2001). "Conceptual Analysis and Reductive Explanation." *Philosophical Review* 110: 315-61.
- Davidson, D. (1980). "Mental Events." *Essays on Action and Events*. Oxford, Clarendon: 207-25.
- Dennett, D. (1987). *The Intentional Stance*. Cambridge, Mass., MIT Press.
- Field, H. (1972). "Theory Change and the Indeterminacy of Reference." *Journal of Philosophy* 70: 462-81.
- Frege, G. (1997). "Sense and Reference." *The Frege Reader*. M. Beaney. Oxford, Blackwell.
- Jackson, F. (1998). *From Ethics to Metaphysics: A Defense of Conceptual Analysis*. Oxford, Oxford University Press.
- (1998). "Reference and Description Revisited." *Philosophical Perspectives* 12: 201-218.
- Kripke, S. (1979). "A Puzzle about Belief." *Meaning and Use*. A. Margalit. Dordrecht, D. Reidel.
- (1972). *Naming and Necessity*. Cambridge, Mass., Harvard University Press.
- Lewis, D. (1994). "Reduction in Mind." *A Companion to the Philosophy of Mind*. S. Guttenplan. Oxford, Basil Blackwell: 412-31.
- (1974). "Radical Interpretation." *Synthese* 17: 331-44.
- (1970). "How to Define Theoretical Terms." *Journal of Philosophy* 67: 427-46.
- (1969). *Convention*. Cambridge, Mass., Harvard University Press.
- Moran, R. (1994). "Interpretation Theory and the First Person." *Philosophical Quarterly* 44: 154-73.
- Peacocke, C. (1992). *A Study of Concepts*. Cambridge, Mass., MIT Press.
- Perry, J. (1979). "The Problem of the Essential Indexical." *Nous* 13: 3-21.
- Putnam, H. (1975). "The Meaning of 'Meaning'." *Minnesota Studies in the Philosophy of Science* 7: 131-93.
- Quine, W. V. O. (1960). *Word and Object*. Cambridge, Mass., MIT Press.
- Root, M. (1978). "Davidson and Social Science." *Truth and Interpretation: Essays in the Philosophy of Donald Davidson*. E. Lepore. Oxford, Basil Blackwell: 165-89.
- Schiffer, S. (1990). "The Mode of Presentation Problem." *Propositional Attitudes: The Role of Content in Logic, Language and Mind*. C. Anderson, J. Owens. Stanford, CSLI Press.
- Schroeter, L. (forthcoming). "The Limits of Conceptual Analysis." *Pacific Philosophical Quarterly*.

sawa who commented on the APA presentation. For helpful discussion and comments, I'd like to thank David Chalmers, Andy Egan, Adam Elga, Martin Davies, Janice Dowell, Frank Jackson, Kirk Ludwig, Peter Menzies, John O'Dea, Philip Pettit, Denis Robinson, François Schroeter, and Daniel Stoljar. Thanks also to two anonymous referees for this journal whose comments led to significant improvements in the paper.