Reference magnetism and the reduction of reference

J.R.G.Williams

(October 8, 2010)

Contents

1 Lewis's interpretationism **2**
1.1 Interpretationism 2
1.2 Reference magnetism and naturalness 4
1.3 Humean simplicity 5

2 Conventionality in metasemantics **6**
2.1 Field on conventionalism 7
2.2 Metametasemantics 10
2.3 Auxiliary apparatus 11

3 Credible reference magnetism **12**
3.1 Response 1: Macronaturalism. 12
3.2 Response 2: Comparative naturalism 13
3.3 Response 3: Parochial eligibility 15

4 Building a credible theory of parochial eligbility **16**
4.1 Two variants of parochialism 17
4.2 Conventionality again 18

Some things, argues Lewis, are just better candidates to be referents than others. Even at the cost of attributing false beliefs, we interpret people as referring to the most interesting kinds in their vicinity. How should this be accounted for? In section 1, I look at Lewis's interpretationism, and the reference magnetism it builds in (not just for 'perfectly natural' properties, but for certain kinds of auxiliary apparatus). In section 2, I draw on (Field, 1975) to argue that what properties are reference magnetic may be an ultimately conventional matter—though in the Lewisian setting, there may be an objectively best conventional choice to make. But Lewis's own account has implausible commitments, so in section 3 I consider variations and alternatives, all of which have problems. In section 4, I look in more detail at eligibility-based interpretationism that do not appeal to naturalness, arguing that there are credible metasemantic theories of this form.

1 Lewis's interpretationism

1.1 Interpretationism

David Lewis was no fan of primitive intentionality. He wanted to explain how the intentional—including mental and linguistic representation —could arise in a fundamentally physical world. He would agree, I think, with Hartry Field:

> there are no "ultimately semantic" facts or properties, i.e., no semantic facts or properties over and above the facts and properties of physics, chemistry, biology, neurophysiology, and those parts of psychology, sociology, and anthropology that can be expressed independently of semantic concepts. (Field, 1975, p.386)

Lewis's *interpretationism* about reference adopts a distinctive reductive strategy: we are told what it takes for a given semantic theory to be 'selected' by the practices of a linguistic community. *What it is* for N to refer to o (in that community's language) is for the selected semantic theory to entail that N refers to o. To give this template content, we need to say what it is for a semantic theory to be 'selected'. Interpretationists favour a two-step answer. First, we specify (in terms of the community's practices) a 'target' pairing of sentences with semantic values. Second, we describe how that pairing, in that situation, picks out a semantic theory.

There are lots of ways to fill in the details. Lewis's own favoured version of the first step pairs sentences with (coarse-grained) truth-conditions; the pairing is established by appeal to conventional regularities in linguistic usage. Thus, if there is a convention among the relevant community of only uttering 'la neige est blanche' when they believe that snow is white, the pairing between that sentence and the set of situations where snow is white is established.[1] A different interpretationist account often associated with Lewis, 'global descriptivism', requires pairing sentences with truth *values* rather than truth *conditions*; perhaps every sentence that is treated by a community as 'platitudinous' is paired with the true; and every sentence that is rejected off-hand by the community is paired with the false.[2]

[1]Notice that Lewis appeals to an intentional relation—belief—in formulating the target pairing. This is in line with his 'headfirst' methodology to the reduction of intention, whereby linguistic intentionality (reference, truth, etc) is reduced inter alia to mental intentionality (belief, desire, etc); a separate story of how the content of attitudes is fixed is promised. See Lewis (1994) for the headfirst/wordfirst contrast, and Lewis (1969, 1975) for an account of the relevant linguistic conventions.

[2]See Lewis (1983, 1984). Lewis is clear there that he is treating the theory as a simplified stalking horse, to illustrate and solve problems with his own more sophisticated account.

The second step is to use the pairing to select a semantic theory. One requirement we could give here is simply this: to be selected, a semantic theory must assign semantic values to sentences that it with the pairings on the target list. (To keep things simple, let's understand 'fit' in the most naive fashion: that the semantic theory predicts a pairing of sentences and contents that exactly matches those that appear on the target list.[3]) One minimal account says that this is necessary and sufficient for selection; more elaborate accounts impose further selectional constraints.

In the early 80's, Lewis began advocating a constraint on the selection of semantic theory over and above fit. He held that some theories were more 'eligible' than others to be selected, because they assigned to lexical items more 'natural' referents. This requires some explanation.

At the time he amended his interpretationism, Lewis had been convinced that for many tasks, throughout philosophy, an appeal to a distinction between 'perfectly natural' properties and merely abundant properties was required. The distinction was more-or-less primitive; perhaps a fan of Armstrongian Universals could explain it in terms of those properties that are necessarily coextensive with a Universal; perhaps we have a choice between positing primitive naturalness and positing a suitably rich contrastive primitive resemblance relation, etc—but for present purposes this won't matter. Lewis also had some specific proposals about what properties had this status. Fundamental physics, he thought, would be our best guide to what the perfectly natural properties of the actual world were.[4] Lewis thought that this account of perfectly natural properties allowed us to make sense of a notion of relative naturalness—or more strongly, degrees of naturalness. The degree of naturalness of a property, he says, would be the minimal length of a definition of that property in terms of the perfectly natural. Thus, for example, if having positive charge and mass of 1kg are each perfectly natural properties, having positive charge *and* unit mass might have a degree of naturalness of 1, where the degrees increase as naturalness decreases). Lewis says that semantic theories are to be graded as more or less *eligible*, depending on the naturalness of the semantic values that they assigned to lexical items—-the more natural (i.e. the shorter the definitional distance to the perfectly natural) the more eligible the theory.[5] Eligibility of semantic theory, Lewis proposes, is a factor to be traded off against fit.

Before looking at this in more detail, it's worth introducing one more concept. Suppose have two semantic theories differing only on the interpretation of one term T, which are equally good as far as fitting the target pairings. It can happen, on Lewis's story, that one is selected and the other is not—this will happen if the candidate referent for T on the first, say, is more natural than the candidate referent for T on the second. Picturesquely, we can say that relatively natural semantic values are 'reference magnets'—all else equal, they'll end up as referents. Notice that reference magnetism as such is not analytically linked to 'naturalness', 'eligibility' and the rest—it is because of the substantive claim that Lewis makes that semantic theory selection should be sensitive to eligibility, that we get the prediction that more natural properties are magnetic in the relevant sense.

Why bother with all this? The motivation for bringing in *something* other than simple fit with the target data is that, without it, the selection of semantic theory is simply too unconstrained. For example, if the only constraint on theory selection is that we generate the right pairing of sentences with truth values, we're open to the observation that if a sensible seeming interpretation T generates the right pairing, we can construct a 'permuted variant' T' which

[3]See Lewis (1975) for discussion of all sorts of sophisticated elaborations of 'fit'—for example, taking pragmatic factors into account.

[4]The canonical statement of Lewis's views here is in his (1983).

[5]Though see also Lewis (1992), where the naturalness of the compositional axioms is appealed to. This will fit naturally into the perspective given below.

generates exactly the same distribution of truth values of sentences. Where the original theory said that 'est blanche' picked out white things, the permuted variant might say that something was in the extension of that predicated iff its image under the permutation ϕ was white. So we overgenerate selected theories; and this is the source of the famous 'inscrutability of reference' problem.[6] Lewis's eligibility constraint speaks to this concern—for even though the original and permuted variant are equally charitable (assign the same truth values) prima facie the permuted variant will be less eligible than the original—it takes slightly longer to spell out in perfectly natural terms. So Lewis hopes that the relative reference magnetism built into his theory will allow us to resist the permutation inscrutability arguments (and all sorts of other related difficulties) and gives us the straightforwardly right verdicts on what natural language terms refer to.

But there's another motivation too: reference magnetism seems a good descriptive fit for how we think about reference. Consider the discovery that simultaneity is relative; that rest and relativistic mass are distinct;[7] or that there are deep tensions within a folk conception of justice. Often, we want to describe these cases as ones where we have been talking about the interesting underlying kinds all along—but had been believing false things about it. It wasn't that we used to report correctly that 'simultaneity is non-relative', and special relativity shifted the subject-matter.[8] Rather, people used to think falsely, of simultaneity, that it was non-relative. We could, no doubt, find some devious interpretation using highly disjunctive properties or context dependency which vindicates much more of old theory. But that's not what we want to do—indeed, it would be sad to have to regard our past selves as talking truly about some uninteresting subject-matter, as opposed to speaking sometimes falsely about the really interesting questions. Reference magnetism, as generated from Lewis's eligibility-inflected interpretationism, can predict the right thing here.

1.2 Reference magnetism and naturalness

The presentation above is deliberately sloppy in a couple of respects, which I'm now going to address. First, I gave Lewis's characterization of degrees of naturalness in terms of 'definitional distance' of properties from the perfectly natural ones. But definitions are given in a certain language, and we don't pick out a language simply by listing a bunch of properties. Sider has suggested we respond to such concerns by extending the natural/non-natural distinction to entities of all categories, rather than just properties—in which case we can envisage a language 'Ontologese' whose every bit of vocabulary stands for something perfectly natural. Whatever the merits of this, I take it that this is not available to Lewis himself. For Lewis, I think, the idea is to focus on definitions within a 'canonical' language, which would include (say) conjunction, negation, unrestricted first-order universal quantification and identity. Perhaps it includes plural quantification and mereological overlap too; and of course the usual variable and punctuational symbols. Let's call these collectively the 'auxiliary apparatus' of the canonical language. Aside from the auxiliary apparatus, the canonical language only contains predicates for perfectly natural properties. We can then say that P is definable in this language if there's an open sentence $\phi(v)$ of the language, such that necessarily, for all x, x has P iff x satisfies $\phi(v)$.

[6]For some of the original literature, see (Jeffrey, 1964; Quine, 1964; Davidson, 1979; Wallace, 1977; Putnam, 1981). See Williams (2009) for a general version of the permutation argument in application to rich languages, and a discussion of why inscrutability is to be avoided.

[7]compare Field (1973)

[8]Correlatively with the relativity of simultaneity is the path-relativity of time—proper time along a trajectory being an invariant notion, by contrast to the time index of the usual coordinate systems that vary under Lorentz transformations.

However, semantic theories don't only assign semantic values to predicates; they also assign them to names, modifiers, operators, and so forth; and we may (indeed, I think we do) want to talk of the relative naturalness of the semantic values associated with these other kinds of lexical items. So we need an extended notion of definition to cover these cases. I'll assume we have one.

If we have the 'canonical language' laid down, and we have some favoured way of measuring the length of the definiens ϕ, then the notion of degree of naturalness can be taken to be the minimal length of such a definition—at least for those terms that have definitions at all. This is one locus for variation in a Lewis-style treat of eligibility, since it's not immediately obvious how lengths are to be measured, nor even what formal structure the 'degrees' will take—will they induce a total or partial ordering? Ordinal or cardinal? Lexiographic or Archimedian? Even if we don't want to commit to a full theory of relative naturalness at this stage, fixing on formal features of the ordering is important. I will assume that the lengths are measured by integer values, so the ordering is total, cardinal and archimedian (and with a natural zero). If you want a toy implementation, suggested by Lewis's writings, imagine that the length of ϕ is determined by counting the number of connectives that are present in ϕ. Notice that relations corresponding to the auxiliary apparatus—perhaps including identity and overlap—will be maximally natural, by this measure, even if they didn't appear on the list of perfectly natural properties and relations. Ultimately, these things will turn out to be 'reference magnets' by the lights of Lewis' theory, just as much as the natural properties are. (If you feel queasy about this, and are prepared to engage in some further meaty metaphysics, you might consider the Siderian alternative).

But interpretationism does not work directly with relative naturalness. Instead *eligibility*—which I'm using as a term for a ranking of semantic theories—is the primary concern. This is a second locus for variation within the Lewisian account—for even assuming that it's only the naturalness of properties that matters, we're being asked to move from a ranking of individual properties, to a ranking of semantic theories that assigns many properties, of various degrees of naturalness. I propose we think of the degree of eligibility of a theory as the sum of the degrees of naturalness of the semantic values it assigns in the lexicon—but notice that this only makes sense because we assumed that the initial comparative naturalness ranking assigned degrees that it makes sense to 'add together'. If we'd instead thought of the degrees as taking a partially-ordered structure, this would not be available to us. We'll come back to this point below.[9]

1.3 Humean simplicity

It's a familiar point that the (syntactic) simplicity of theories in general depends on the language in which they're formulated. To give Lewis's example: if we have a primitive predicate F, that expresses 'being such that T holds', then the single axiom $\exists x Fx$ will give us a maximally simple informational equivalent of T. But we want to compare theories for simplicity even if they

[9] I've been writing as if the semantic theory proceeded by assigning semantic values to lexical items, and this seems natural if we think of something like a Lewis (1970) general semantics rather than, for example, a Davidsonian T-theoretic semantics (Larson & Ludlow, 1993). But even a general semantics can't do everything by the assignment of semantic values—Lewis's theory included the compositional axiom of function-application, and later advocated syncategoramic axioms governing lambda-operators. And it is arguable, I think, that even those sympathetic to possible worlds semantics should take a 'semantic theory' not simply to be an assignment function mapping expressions to semantic values, but rather an axiomatic theory that specifies such a function (see (Heim & Kratzer, 1998) for one version of what this would look like). There are natural adaptions of the above ideas to this setting—what takes the place of relative naturalness will be the length of the axiom governing a particular lexical item when formulated in the canonical language; and we again determine overall eligibility by adding this up.

differ in vocabulary. This had generated a puzzle for Lewis in connection with his Humean theory of laws of nature—he wanted to pick out the laws in terms of the consequences of a 'best' (optimally simple and informative) axiomatic theory. But if by switching languages we can maximize simplicity without sacrificing informativeness, then the simplicity component of the Humean theory loses its bite. In reaction, Lewis proposed that the relevant notion of simplicity operates on presentations of the theories concerned in a 'canonical' language, built out of the perfectly natural predicates—indeed, exactly the canonical language relevant to assessing relative naturalness. It is syntactic complexity in this privileged representation that matters for Humean system selection—and the introduction of artificial predicates F as above is neither here nor there. Let's call the notion of simplicity so characterized 'Humean'.

We can inquire into the Humean simplicity of all sorts of theories—in particular, semantic theories. At this point, it's natural to draw connections as follows: take an axiomatic presentation of semantic theory *in the canonical language*. The complexity of the axiom assigning a semantic value to an expression *in this presentation*, will, modulo some constant factor, be given by the complexity of the definition of that semantic value in canonical terms (what we called earlier the degree of naturalness of that semantic value). Adding the complexities of axioms together gives the overall Humean complexity of the semantic theory we start with; but it also (modulo some constant) gives the overall eligibility of the theory. So Humean simplicity/complexity and eligibility/(in)eligiblity measure the same thing.

Now, this line of thought is resistable—we might want to develop these two, obviously related ideas in different directions. But I can't see why we'd want to. If anything, thinking of overall simplicity of theory seems a good way of seeing areas where the account of eligibility needs elaboration (for example, the formulation of the compositional axioms is factored into simplicity, and certainly should be accounted for in eligibility—we don't want crazy interpretations of compositional axioms messing things up for us (cf. Lewis, 1992)). So, my preferred take on Lewis's eligibility-based interpretationism is to formulate it directly in terms of Humean simplicity; the proposal being that the selected semantic theory is the one that optimizes *simplicity* and fit with the target data.

2 Conventionality in metasemantics

I argued earlier that one attractive feature of eligiblity-based interpretationism is that it correctly predicted certain familiar reference-magnetic phenomena, as in the cases of mass and simultaneity. You might wonder about how good an explanation it provides, however—what eligibility seems to do is identify a bunch of properties as the 'reference magnets'— but isn't what we're after an explanation of *in virtue of what* magnetism occurs?

It's not hard to find sympathy with this reaction. Why would the mere fact that some property is metaphysical fundamental mean that our language is more likely to refer to it over rivals—especially when no speaker has a clue about these fundamental properties are? What the words of the ancient Romans picked out, on this account, is fixed by the relation of the referents of their words to mass, charge, charm, and/or whatever else physics eventually throws back at us. Even weirder is Lewis's mixed approach, where there is ultimately no unifying characteristic of the reference magnets—as well as the perfectly natural, auxiliary resources such as quantifiers and connectives magnetize. So we again ask the question: *why* are these entities magnetic? In the case of perfectly natural properties, we have some feature to work with—metaphysical fundamentality— but we have no clue why that feature should be relevant. In the case of the others, it's not even clear what the starting point of an explanation would be.

Against this puzzlement, there's a flat-footed response. We judge the success of a theory

primarily on the basis of successful predictions. If a metasemantic theory gets the right results—if in application to arbitrary languages, it gives plausible results about what refers to what; if it has general theoretical virtues and beats its competitors on this basis—-then we have reason to believe in it. The Lewisian interpretationist will say that their particular story wins out on these grounds. If the account then entails that a certain range of entities are reference magnetic, so be it— this is a philosophical discovery, not a mysterious posit on which the credibility of the account rests.

Both sides can, for the sake of argument, agree that the theory in question predicts the data well, and that if that were the only issue at stake, we should endorse it. But one side sees an unsatisfied explanatory debt; the other side thinks there's nothing there to explain. Such debates are hard to adjudicate. I think the best way to get a grip on this debate is to consider the relation between the eligibility-based interpretationism that Lewis advocates, and some closely related theories that say different things about which properties reference magnetize. If there's a satisfying answer to those demanding further explanation, it'll emerge in the comparison of these theories.

2.1 Field on conventionalism

Suppose Sensible Sandy advocates Lewisian reference-magnetism. Crazy Cate advocates a different view: retaining the form of eligibility-based interpretationism, but switching out Lewis's favoured canonical language, in favour of a 'permuted' alternative. In particular, while Sandy evaluates the relative naturalness of an entity in terms of its definability from within a language where each of the predicates stand for perfectly natural properties, Cate evaluates relative naturalness via a canonical language where each predicate stands for the ϕ-image of a natural property for some permutation of the universe ϕ—where Q is the ϕ-image of P if and only if necessarily, x instantiates Q iff $\phi(x)$ instantiates P. Eligibility so-construed would favour an interpretation of English under which singular terms determinately refer to the ϕ^{-1} images of what we would standardly take to be their referents. And this kind of holistic permutation of the referents of lexical items will leave the truth-conditions of whole sentences unchanged.[10] So Sandy and Cate disagree about the extension of the reference relation; and disagree about the metasemantics that fixes reference; but agree about in what circumstances sentences are true.

This *kind* of challenge to metasemantic theories is discussed by Field (1975). In the context of a causal theory of reference, Field points out that rather than building a metasemantic theory that appeals to causation, we could build one that appeals instead to the causation*—something that relates e to f iff e is the ϕ-image of some event that causes f. And from this we can characterize a word-world relation, reference*, which when fitted into the standard Tarskian framework would characterize the very same truth-conditions as genuine reference does. Analogues of Sandy and Cate could take one or other package. The challenge is: what justifies us in thinking Sandy is sensible and Cate crazy? Why should we, as theorists, endorse one over the other?

Field's answer is interesting. He notes, first, that reference (characterized via causation) and reference* (characterized via causation*) are two perfectly genuine relations between our words and the world. He thinks that we do in fact use reference, rather than reference*, in our semantic theorizing (this follows from what Field calls the 'conformity requirement' (p.379)—that semantic relations we focus on conforms, more or less, to our ordinary usage of semantic terms). So we have available a flat-footed response to the query just mentioned—the metasemantic theory is attempting to spell out what grounds the facts, inter alia, about *reference*—and

[10] Again, see the references cited in the earlier footnote, and for my own take, (Williams, 2009).

to appeal to causation* and reference* would be to change the subject. But, says Field, to leave it there would be to miss a deeper point. In his view, the fact that semantics itself is formulated in terms of reference rather than reference* is at root a convention: we could have systematically theorized about content via reference*, for example; and nothing of real substance to the semantic enterprise would be lost.

To understand this claim, we need to have an idea of what the 'real substance' of semantic theorizing is. Let us focus on the *deployments* of semantic properties—their role in wider theory. If we're asked what the *point* is of thinking about semantic properties at all, then we'll naturally look to these deployments. Field emphasizes elsewhere (Field (1978)—see in particular the postscripts in Field (2001, p.72)) the role of content attributions in psychological explanation (construed as the content of mentalese sentences). Suppose—what is certainly arguable—that assignments of *subsentential* content plays no explanatory role in this wider theory— the only thing that matters is the overall content of whole thoughts.[11] If that's the case, then *what matters* if semantic theorizing is to do its job, is that it deliver the right results about the truth-conditions of whole sentences. Accordingly, let us say with Field (1975, p.377) that a semantic theory is *holistically adequate* if it generates reasonable results about the truth-conditions of whole sentences—ex hypothesi, any holistically adequate theory will deliver ok results so far as wider theory is concerned. Field's 'Requirement A' on a semantic theory is that it be holistically adequate.

We've already noted that both Sandy and Cate's accounts agree on truth-conditions of whole sentences, and hence they both meet requirement *A*. There's another thing we can also say—the operative notions they use to characterize the truth-conditions (reference and reference*, respectively) can be reductively characterized in terms that do not presuppose primitively intentional facts. That a semantic theory generate truth-conditions out of a theory that can be reductively characterized in some way is Field's 'Requirement B'.

In Field's view, requirements *A* and *B* have an entirely different status from the conformity requirement. Proposing a semantic theory that violated either *A* and *B* would be propose a theory that didn't do the job we were relying on it to do; or one that only did so by appealing to resources to which it wasn't entitled. Sandy and Cate's accounts are on a par at this level. But noting that Sandy's theory in addition meets the conformity requirement is just to note that she has chosen to discharge the task in the same way as we do; that in no way shows that Cate's account is *worse off* from an independent point of view. This is why the comparison to conventions seem appropriate. Our road traffic laws aim to enable people to get from one place to another with minimal accidents. Enforcing driving on the left is one way to achieve this; so is enforcement of driving on the right. The UK opts for the former option; but either would achieve the objective, and choosing between them is paradigmatically conventional.

Turning back to Lewis's interpretationism, the dialectic that Field runs for the causal theorist can be run all over again. Rather than causation and causation*, we can talk of naturalness and naturalness*, and the reference and reference* relations that are thereby induced. If Field was right about the scope of the theoretical role of semantics, then ultimately our choice to focus on one or the other may be conventional. But building a theory on naturalness* is perverse—an unsmooth way of satisfying a theoretical need.[12]

[11] In effect, Williams (2009) makes a case that *stability* of subsentential reference makes a contribution to wider theory—in particular to the epistemology of inference. This wouldn't rule out the reference* relation, on which what plays the reference-role is determinate and stable, albeit weird.

[12] We should be careful to distinguish two elements here: a metasemantic theory may say that what makes something the meaning-fixing theory is in part the theoretical virtues (fit, simplicity, eligibility) displayed by the semantic theory in question. But we may look at the metasemantic theory itself, and evaluate whether the account of reference it gives—in particular, the treatment of "eligibility" it includes—is simple, plausible, elegant and so

So there is a Fieldian case to be made that the focus on natural properties, in particular, in Lewis' metasemantics, is conventional (though none the worse for that). It would be no additional hardship, I think, if we acknowledged similar conventionality in the choice of auxiliary resources. The choice of whether to include conjunction and negation, or the Sheffer stroke; whether to include parthood and overlap or only one or the other, in the canonical language, is very plausibly seen as conventional in the above sense. This suggests a direct response to the argument to explain in virtue of what the reference magnets magnetize—the answer being that there is *no* deep reason for them to be reference magnets. To seek a further explanation is like seeking an explanation of *in virtue of what* the right hand side of the road is the one to be driven on. That they are the magnetic in this sense is an artefact of an essentially arbitrary decision to focus on one among a slew of ways of discharging a certain theoretical task.

Acknowledging conventionality so construed doesn't prevent us ranking some conventional choices as better than some others. Consider again road-traffic conventions. It's perfectly conceivable that one choice is all things considered a better— maybe the right-handedness of the majority population makes right-sided convention marginally safer, for example. Likewise, a pattern in which the side of the road to drive on switched at midday is a possible convention that if implemented would do the required coordinating job—but it's clearly a perversely complicated choice, and radically less safe (one stopped watch causing chaos). Nothing in Field's point stops us from noting that achieving via appeal to reference* rather than reference is perversely complicated (as Field notes (p.378), reference is pretty clearly *simpler*). So we see that there are at least two dimensions for evaluating a candidate for playing some theoretical role. The first is whether it does its job. The second is whether it does that job smoothly and elegantly—in the optimal way. These two kinds of evaluations are independent of the factual issue of which candidate we in fact use for the task in question–which proposal meets the conformity requirement.

Let's call conventionality where there's *little to choose* between rival conventions *deeply conventional*. While there might be some small advantages to a left-side over right-sided driving convention, I take it that they're roughly on a par on grounds of simplicity, safety etc. So the choice between them is deeply conventional. On the other hand, the driving convention which switched sides at midday is clearly inferior, because far less safe. So the choice between a uniform driving convention and a switching one is only a shallowly conventional choice.

Field argued that conventionality is an unexciting doctrine (in particular, it doesn't provide the basis for arguing against a realistic view of reference, or teach us that we need to reconstruct semantic theory in non-referential terms). I don't think it's for that reason obvious or unsurprising. It's certainly not obvious, since it's not clear that only the truth conditions of whole sentences matter for wider theory—and if subsentential content does matter, then the requirement of 'holistic adequacy' can be strengthened in ways that cut down on the range of conventionality. How surprising it is depends on how deep the conventionality runs, in the sense just specified. If the conventionality is shallow, and there's something *clearly* superior about Sandy's views as against Cate's, then that's one thing. If the conventionality is deep, and there's really no very good objective reason to think about semantics in terms of reference in particular, then the theses really is radical. Further, identifying the ways in which a naturalness-based theory is better than a naturalness* based account (if there are any) will illuminate why the particular entities that Lewis focuses upon are appropriately treated as reference magnets. So it is to this we now turn.

on. It's this latter, second-order evaluation that we're appealing to.

2.2 Metametasemantics

We've already mentioned the idea that reference seems a much simpler way to theorize about semantics compared to reference*. Whether this is an *objective* respect in which a naturalness-based metasemantics is better than a naturalness*-based one is open to question. A very minimal way of construing the point is that *for creatures like us*, it's pretty easy to formulate a reference-based semantics, and would take considerable effort to rewrite textbooks in a way that invokes permuted semantic values. But perhaps creatures with other kinds of internal engineering would find reference* the easier notion to work with. Tying respects of betterness to such contingencies would, I think, still leave us with a pretty deep kind of conventionality.

But it's not clear that the appeal to simplicity needs to be understood in such a relativized fashion. Humean simplicity, for example, was characterized in terms that did not appeal to what creatures like us find easy to work with. It aims to specify *how objectively complex* a theory is, where the standards are fixed by the fundamental structure of the world. Humean simplicity has a direct application *to our philosophical theorizing*. To the extent that we have to talk about ϕ-images of natural properties rather than those properties themselves in giving our philosophical theory, we sacrifice Humean simplicity in our account.

But there are two other ways in which appeal to Humean simplicity could speak in favour of a naturalness-based metasemantics. First, on the permuted metasemantics, the properties assigned to object language predicates (i.e. the properties referred* to by object-language predicates) are more complex than on the Lewisian metasemantics. So on this interpretation the *agent's* theories about the worlds will be formulated in needlessly complex terms. When they utter "the ball fell because it was pushed off the edge", what they'd be depicted as saying (i.e. saying*) would be that the ϕ-image of the ball was ϕ-pushed. Their explanations and theories are worse than they would be on the interpretations delivered by Lewisian metasemantics. So this choice of metasemantics gives rise to uncharitable interpretations of agents.[13]

Second, a metasemantics like Lewis's tells us to select the semantic theory that optimizes fit and simplicity; whereas the permuted metasemantics will tell us to optimize fit and some other property. Epistemologically, to figure out the correct theory of reference*, we can't look to the simplest theory of the data—so semantic theorizing in Cate's favoured style will be epistemologically discontinuous with theorizing more generally.

However, these explanations are only good (and report a kind of objective betterness) to the extent that the simplicity verdicts on which they rely are good (and make for objective betterness). One might be sceptical that Humean simplicity, characterized as complexity in the canonical language, can play this role (see, for example, Loewer (2007) for an expression of scepticism). Is the general theory of relativity a simpler theory than the straightforward differential equations of population ecology, just in virtue of the latter being longer than the former when spelt out in a canonical language that is constructed from the metaphysically fundamental? In general, the account of Humean simplicity looked reasonable when (a) we were interested in comparing the theories of fundamental physics; (b) we presupposed that the 'perfectly natural terms' are in fact drawn from the microphysics. It's far less obvious that this treatment is appropriate when we are interested in evaluating the relative simplicity of special sciences. So for example, it is indeed attractive to view interpretationism as the projection of the best (inc. optimally simple) semantic theory onto the world. But it seems surprising, prima facie, to theorize about the standards of simplicity appropriate to semantics as a special science by appeal, inter alia, to microphysics.

[13]Note that 'charity' is often talked about in connection to radical interpretation as a principle of i nterpretation-selection—something that plays the role given above to fit and eligibility. I'm appealing to it here at a different level—metametasemantic.

I'll look in the next section at accounts that vary the canonical language to take account of the points just raised; but for now I want to sketch a way of defending what we've seen so far. There are really two potential challenges here: one is that the account of simplicity we rely on gets things extensionally wrong (delivers the wrong comparative simplicity verdicts, e.g. between ecology and general relativity)—and another is that its starting point is ad hoc and unprincipled, at least when we construe it as an account of simplicity in general. On the first point, there are things to say: for example, its not clear that cross-topic comparisons of simplicity should bear much weight—it's not as though we're typically engaged in theory-choice between a biological theory and a physical one. If we suppose that biological theories involve working primitives of roughly the same eligibility (in Lewis's sense), then the fact that both involve long definitional chains to the microphysical should 'cancel out', and we can expect sensible verdicts about which is the simplest.

What of the charge of ad hocery in using the canonical language as the starting point? To take this on, I think we need to appeal to something that has been argued for on independent grounds by Ted Sider. Sider argues that our most general epistemic ambition is not merely to believe truly, but to believe truly *in the right way*—-where the 'right way' involves an isomorphism between the structure of concepts in our thoughts, and the structure of the world—that each concept we use picks out something perfectly natural. If Sider is right that the natural is normative in this way, then we have a response to the charge that it's ad hoc or unprincipled to rely on a language formulated in perfectly natural terms, to evaluate simplicity. What it amounts to is the proposal that we evaluate theories with respect to simplicity when formulated in the *best* way—which on Siderian grounds we take to be their formulation in the perfectly natural language. We rely on the evaluations of syntactic simplicity our ideal selves would make—and that seems as principled as one could wish.

2.3 Auxiliary apparatus

The response just suggested assumes that the only reference magnetism we had to explain was that of the perfectly natural properties. But in Lewis's own account (I claimed), we need to deal with the auxiliary resources included in the canonical language—quantification, connectives, overlap, etc. The first part of the defence of Lewisian metasemantics just sketched (the appeal to Humean simplicity) can be carried across unchanged. However, the second stage (the defence of the characterization of Humean simplicity) is on shakier ground. If the natural/non-natural distinction is restricted to properties, then there's nothing in the idea that the natural is normative that tells us that thinking with an unrestricted quantifier is better than using some restricted alternative.

So we face a challenge analogous to the one above: rather than considering a metasemantics formulated in terms of ϕ-images of natural properties, we could consider a metasemantics formulated with quantifier restricted to the some countable subset of the universal domain (a la Putnam's use of the downward Loweinheim-Skolem theorem), plus natural properties and the usual connectives. We could try to object that the theory we endorse, and the interpretation of agents we're give, are needless complex. That presupposes a standard for evaluating the simplicity of theories that doesn't itself use the Skolemized quantifier. And so we trace the rationale for the universal quantifier being a reference magnet, to the claim that we should evaluate theories for simplicity inter alia using a universal quantifier rather than a Skolemized quantifier. But why should that be? Why should theories formulated with restricted quantifiers count as ipso facto more complex?

There are two important points to recall at this point. One is that Field's own presentation of the underlying issue of conventionality in metasemantics is *conditional*. If two metasemantic

theories give the same results, on the issues that matter to wider theory, then the choice between them is ultimately conventional—at which point other dimensions of evaluation become pertinent (e.g. which is the better/smoother way of discharging the task). A clean case of this can be built with the permuted metasemantic bases, on the assumption that what matters for broader theory is sentential content. But in the case of restricted quantifiers, the choice between Lewisian metasemantics and its Skolemized alternative will probably not be conventional in this sense, since they will attribute different contents to quantified claims. It's one thing to desire that there be, unrestrictedly, no suffering. It's quite another to desiring that there be no suffering in R (for some restriction R). Likewise, a theory that says that everything is physical, is quite different from one that says everything that is R is physical. Being able to think, theorize, and desire unrestrictedly (that is, with the *contents* that arise from using unrestricted quantifiers) is important to us—and this, I think, is sufficient justification for adopting a metasemantics which makes it possible, given that one is on offer.

For some auxiliary devices, there might be nothing comparable to say. For example: exactly which range of connectives should the canonical language feature? Should it be formulated in terms of overlap, or parthood, or proper parthood? Insofar as the choices make no difference to the truth-conditions of sentences attributed, one might think that here we have a really deep kind of conventionality—something where there's just nothing to say about why this or that relation in particular is reference-magnetic. But this doesn't seem counterintuitive at all; indeed, a deflationary attitude to reference-magnetism in this particular area seems quite appropriate. And it is perfectly possible to combine this with a non-deflationary attitude to the reference magnetism of perfectly natural properties, unrestricted quantification, and other things.

3 Credible reference magnetism

The above discussion presupposes that we get sensible results out of a Lewisian metasemantics, on which eligibility is analyzed ultimately in terms of definitional distance from the perfectly natural—which in Lewis's case, meant microphysical properties. But many feel that we are entitled to no such assumption. I don't really have a clue what a 'definition' of the ordinary subject matter of thought and talk—shoes and string and sealing wax—would be, if the definiens is to be drawn from microphysics. Further, I think that there are specific reasons to be worried that the account gives the wrong results—see Williams (2007a).

I won't argue directly against the Lewisian proposal here—so I'll leave it open for others to make the case that ordinary notions are finitely definable in the required way, and that the arguments that it leads to trouble can be blocked (perhaps by being appropriately subtle about some of the loci of variation in the account—the relation between perfect naturalness and relative naturalness, and relative naturalness and eligibility). Instead, I will work with the assumption that the Lewisian proposal as originally envisaged fails, and examine what prospects remain for an eligiblity-based interpretationism.

I'll look at three proposals: two that revise the overall metaphysics, and one that keeps the metaphysical framework intact but drops the connection between eligibility and naturalness.

3.1 Response 1: Macronaturalism.

The troubles for Lewisian eligibility, it might be thought, do not originate from the theory of eligibility per se, nor in its relation to perfect naturalness. The worries stems from the fact that Lewis commits to the view that (in the actual world) the perfectly natural properties are to be found in microphysics, and not in the 'macroworld' in which we operate. But, one might

argue, wherever we find law-like connections; wherever we find genuine objective similarity; and wherever we find causation, we should believe that we're working with perfectly natural properties. And, it may be argued, we encounter such phenomena in the macroworld of geology, biology and ecology as much as the microworld of theoretical physics (Schaffer, 2004). The perfectly natural properties will be sparse but not ultra-sparse.

Whether eligibility-based interpretationism remains reductive in this setting is open to question—if we're allowing in the equivalent of biological and ecological Universals, what about psychological ones—beliefs and desires? Indeed, what about the special science of semantics? One might think that a principled version of this macroworld picture should include perfectly natural relations corresponding to intentional verbs, or even reference itself.

Even if we can take it that the vocabulary of the special sciences is available to us, it's not clear how we'd go about defining terms for artifactual kinds, nor the variety of verbs we use in everyday life (terms for thick ethical or aesthetic concepts, for example). So even if the range of resources we have available as a definitional base isn't as recherche as on the Lewisian proposal, the definitional ambition is still grandly ambitious. Of course, the reductive achievement in prospect, and the definitional ambition, play off one another: the more sparse a macroworld we buy into, the more reductive the final proposal, but the bigger the definitional task we set ourselves.

The macroworld view of the distribution of perfectly natural properties raises some questions about what the wider role of natural properties is to be. One idea that is prominent among Lewisians is that natural properties should enjoy a certain kind of modal independence. Chalmers' dualism (Chalmers, 1996) is a natural illustration of the sort of thing we might expect: on this kind of view, if some physical thing is conscious, it's possible for there to be a physical duplicate of it that is not a duplicate simpliciter, because it lacks consciousness. One question for the macronaturalist is whether something similar goes for their properties. Is it possible to have physical duplicates that are not chemical or biological duplicates, for example? If not, aren't we committed to some objectionable 'necessary connections between distinct properties'? Again, if special science kinds (for example) are perfectly natural, then won't the vagueness and indeterminacy of (e.g.) biological kinds give us 'vagueness in the world'?[14]

But the macronaturalistic view is not something I'd be comfortable on relying on: we should explore whether there's some more localized and modular response to our problems.

3.2 Response 2: Comparative naturalism

A second response to worries about definability from microphysics is that Lewis went wrong (or that I went wrong in interpreting Lewis) by attempting to reduce relative naturalness to perfect naturalness via 'lengths of definitions' from some canonical language. Maybe we should stick with relative naturalness itself as primitive. After all, the motivating cases for this distinction are examples that involve comparative judgements very distant from what Lewis regards as perfectly natural: that green is more natural than grue; that being the image under a permutation of something human is less natural than being human; that artefactual kinds are less natural than biological kinds—and so forth. If comparative naturalness is rock-bottom, then an addition bonus (one might think) is that we can be agnostic over whether there is a layer of 'maximally natural' properties in the first place opens up—perhaps there are simply more and more natural properties ad infinitum (compare Schaffer, 2003; Langton & Lewis, 1998).

Just as with macronaturalism, we need to consider how primitive comparative naturalism

[14]For some exploration of life with a primitive macroworld, see Williams (2008b, 2007b). For vagueness in the world, see Williams (2008a); Barnes & Williams (2010).

integrates into wider theory. For example, it might sound attractive to countenance the possibility of ever-more-natural properties; but if the theoretical deployments of naturalness appeal to the all-or-nothing concept, then it's not clear that comparative naturalness will be an adequate replacement. As an illustration: Lewis's original theory of duplication and intrinsicality made appeal to the sharing of perfectly natural properties (cf. Lewis, 1983, 1986). The theory of Langton & Lewis (1998), which proved far more problematic, is one exactly designed to liberate the analysis of intrinsicality from appeal to a layer of maximally/perfectly natural properties. That it runs into worries that do not face the original account illustrates the damage such shifts in resources can bring.

There are more local concerns about the adequacy of an appeal to primitive comparative naturalness in connection to metasemantics. Recall that earlier we emphasized the distinction between relative naturalness (of properties) and relative eligibility (of whole theories). Let's think of a toy case: an object language that has the syntax of first-order logic, with only three non-logical terms, the predicates 'rock', 'tree' and 'human'. To cut down on complexity, look at semantic theories that differ only over their assignments to the non-logical predicates. Compare three candidate assignments to 'rock', 'tree' and 'human', respectively: $A = (A_1, A_2, A_3)$, $B = (B_1, B_2, B_3)$ and $C = (C_1, C_2, C_3)$. We may suppose, writing $>$ for 'more natural than', we have: $A_1 > B_1 > C_1$, $B_2 > C_2 > A_2$ and $C_3 > A_3 > B_3$. This component-wise ranking gives us three different induced rankings of the overall assigned, which display a cyclic Condorcet-structure: A beats B twice, B beats C twice, and C beats A twice. In that kind of situation, how are we to rank the candidate assignments for relative eligibility?[15]

We earlier pointed out that one question for the theorist of comparative naturalness faced was over the formal characteristics of the ordering. Partial or total? Cardinal or ordinal? Lexiographic or Archimedian? A zero element or not? On some ways of answering this question, there will be natural ways of sorting out Condorcet-style troubles—for example, in the Lewisian account that (I took it) delivered integral degrees of naturalness, we could simply look at the sum total of the various degrees. But that kind of structure seemed reasonable because we were analyzing comparative naturalness in terms of a metrical notion—lengths of definitions. In the current context, there's no reason yet to think that summing 'degrees' of naturalness makes sense (particularly if they're partially ordered, or have no natural zero). And of course, the richer the structure we're forced to posit, the more costly it is to take it as explanatorily basic. The moral from these discussions is that we need to keep sharply in view that the resource that interpretationism needs is a ranking of whole theories, not individual properties. Without a story of how to get from the latter to the former, we have no theory at all.

If I were tempted by this direction at all, I'd take a different route. Consider the following analogy: truth is an aim of belief. But as finite fallible creatures, we often believe in ways that aren't simply classifiable as simply achieving that aim or vitiating it. In particular, we adopt *degrees of belief* in propositions. Evaluation of how well we achieve the aim must then be more subtle—we should talk about the *degree of accuracy* a given partial belief state has, given the world (for the notion of accuracy as a gradational alethic norm, see (Joyce, 1998, 2009)). Likewise, following Sider, there is an all-or-nothing aim for belief: to believe using concepts that perfectly match the objective structure of the world. But again, as finite creatures operating at a relatively macro level, we don't perfectly achieve this aim, and the relevant appraisals concern only how close we get to achieving it. This is what eligibility (and perhaps Humean simplicity) tries to capture. Now, perhaps no reduction of these evaluative concepts of 'closeness to

[15]Following up the analogy to voting paradoxes, it's worth asking whether a version of Arrow's theorem lurks in the vicinity. Pareto and non-dictatorship constraints are certainly plausible. However, it's not clear to me whether there's an interpretation of the Universality and Independence axioms that allows us to invoke the theorem in this setting.

achieving the goal of belief' are available. If so, there's motivation for taking Humean simplicity/eligibility of (structured) proposition as primitive, but no reason to think that we need to take as primitive comparative naturalness of properties, which was introduced as part of a failed strategy for *reducing* eligibility. The eligibility of a semantic theory is then measured by the eligiblity of the structured proposition expressing the semantic theory (perhaps Ramsifying out the semantic relations). I doubt primitive eligibility will win many friends, but it seems to me a better option than primitive comparative naturalness.[16]

3.3 Response 3: Parochial eligibility

The final option I will consider involves dropping the link between naturalness and eligibility, but retaining the rest of the structure of eligibility-based interpretationism. The story developed earlier can be seen as giving a programmatic specification of degrees of eligibility. We input some 'canonical language' L. We obtain an ordering of properties (and other entities) by minimum length of definitions in L—and by summing we thereby obtain what we might call 'degrees of eligibility$_L$' attaching to sets of entities/whole theories. Degrees of eligibility$_L$ are then traded off against fit to select the meaning-fixing theory of interpretationism. So each choice of L 'projects' certain account of reference—reference$_L$. For some L, reference$_L$ is clearly not reference itself.

There are many candidates for L. L could be, for example, English. After all, the metalanguage of semantic textbooks tends to be a natural language (suitably supplemented by technical vocabulary)—not some artificially restricted language drawn from metaphysics or other special sciences. If we start with English, we have something with the recognizable pattern of eligibility-based interpretationism, but the base from which eligibility is determined is parochial rather than metaphysical.

There's clearly something parochial and deflationary about this proposal, and it is instructive to compare it to the disquotational 'analysis' of reference Field (1972) famously argues against. Field presented the view as a kind of listiform reduction: x refers to y iff either x is 'duck' and y is duckhood, or x is 'rabbit' and y is rabbithood, etc. Many think that there's something dodgy about this *style* of analysis—but whatever we say about that, there are many other specific troubles to worry about. For example, it's not *general*—it analyzes reference only for English, not for French let alone merely possible languages, or future versions of English with novel vocabulary. It seems to get the modal profile of reference even in English wrong—had our usage of 'rabbit' and 'duck' been interchanged, their referents would have been switched; yet construing the above as an identification, there seems no room for this. Finally, the 'analysis' appears to be circular, insofar as we use semantic terminology (*refers*, *says*) in disquotationally characterizing the referent of English 'refers' and 'says'.[17]

At first glance, the parochial treatment of eligibility shares both the listiform character, and some versions of the specific vices. What does it take for a property to be an eligiblity-maker? Either that it is rabbithood, or duckhood, or redness, or... and so forth using all the English predicates. That these happen to be the properties picked out by *English* predicates doesn't

[16] I've been talking as if comparative naturalness would be treated as a metaphysical primitive, as perfectly naturalness arguably is by Lewis. But on the view just sketched, it's not clear whether that's motivated. Comparative accuracy need not have the same status as truth; and comparative eligibility need not have the same status as perfect naturalness. Indeed, one might argue that comparative accuracy and eligibility should be treated as evaluative concepts, to be handled in whatever way your favoured theory of normativity suggests. This may, perhaps, mess up the reductive ambitions of eligibility-based interpretationism, but nevertheless has some independent interest.

[17] I'm not meaning to suggest these are criticisms of disquotationalism about reference, as advocated by many people these days (including Field (1994)). I take those theories to have a quite different theoretic ambitions and resources—for example in the prominent role allocated to translation.

seem a unifying characteristic, so the treatment of eligibility is essentially listiform. To some extent it does better on the specific vices. It certainly does better on the grounds of modal profile and generality. If the usage patterns of 'duck' and 'rabbit' were interchanged, then our parochial eligibility-based theory would predict that 'duck' refers to rabbits and 'rabbit' to ducks. The properties picked out by English terms would be the eligibility-makers, meaning that duckhood and rabbithood would be in the front line for being referred to; but whether or not they are referred to depends on whether they fit with usage. Likewise, the theory applies, and can be expected to give at least approximately reasonable results, to natural languages other than English. But as it stands the account has the same trouble with circularity as with the straightforward deflationary proposal. For example, that the comparative naturalness of a property (and so eligibility) is determined in part from the semantic/intentionality vocabulary of English (referring, being true, being believed etc), the reductiveness of the account is vitiated. Furthermore, when we dig a little deeper, we see that all is not well with the account's generality. Consider the following counterfactual situation (which may or may not be realized in actual natural languages): a community uses a colour predicate 'G' approximately as we use green, but systematically favour applying it to bluer shades than we do; and are unwilling to apply it to some shades at the yellowish end. The obvious description of the situation is that they have a colour term that picks out a colour property that is closely related to green, but with slightly shifted boundaries. On the other hand, describing the shifted boundaries will be tricky in English—if it's possible at all, presumably the description will be pretty complex. But interpreting 'G' as green, though it doesn't fit as well, would be maximally eligible, if eligibility is measured (as we're contemplating) relative to those terms that happen to occur in English. This seems to get things extensionally wrong—the current proposal builds in a kind of *imperialism* about distinctions carved by primitive predicates in English, allowing those to override the distinctions suggested by usage. (The naturalness-based account of eligibility promised to be free of such biases, since it provided an *independent* standard of assessment.)

A related problem for generality is that the current account would not be *resource sensitive*. Consider the Ectoplasmians: a population of language-users living in a world of a completely different constitution from our own. We have no words in English for the fundamental properties that structure their environment (though I'll use 'ectoplasm' as a placeholder). And though some concepts may be recognizable, many of the topics of their speech concern things that we have not a hope of characterizing in English. Naturalness-based eligibility, if it worked at all, would work for this case in particular, since it builds in a sensitivity to the local resources on offer: if the Ectoplasmians appear to be talking about some property P definable from ectoplasmic natural properties, then P would thereby be assigned a degree of naturalness, and the machinery churns away. It would be optimistic, to say the least, to think that the subject matter of all possible words in all possible worlds is expressible in English.

4 Building a credible theory of parochial eligbility

Our survey of eligiblity-based interpretationisms has been rather negative, thus far. The Lewisian proposal, with its microphysical foundations, is objectionable. The optimists might think macronaturalism comes to the rescue, but given the metaphysical commitments entailed, we've lost a lot of our audience. Appealing to primitive comparative naturalness raises as many questions as it answers—and personally I see it as hardly less committal than macronaturalism. But the kind of parochial theory just discussed seems beset with objections. Nevertheless, I think a credible theory can be extracted.

4.1 Two variants of parochialism

It is with the parochial approach that there is most room for maneuver. The problems we saw arose because of the specific parochial proposal we've considered. And that leaves open the prospect of some more nuanced account of this kind. For example, the fan of macronaturalism, reluctantly convinced that only microphysics provides perfectly natural properties, might construct an ersatz version of macronaturalism. The eligibility-makers, says she, are those expressed by the vocabulary of final physics, chemistry, and biology, deployed as a listiform analysis of eligibility.[18] Such a theory would have a mix of previously noted virtues and vices. The particular biases of basing an account on English would diminish, since presumably the physical, chemical and biological properties are neutral ground. The circularity worry is gone; but the ambitious definitional programme macronaturalism required is once again needed.

This account has the capacity to be somewhat more resource-sensitive than the straightforward parochial proposal—some varieties of Ectoplasmians might live in a world with identifiable physics and chemistry, albeit configuring properties different from our own. And this kind of parochial theory might have those properties be eligibility-makers. But it's still not as resource-sensitive as Lewis's proposal, where there's no need to appeal to structures of theories analogous to the ones we find useful in our environment. (I suggest that in such cases one should simply deny that the language-use of the Ectoplasmians should be understood via reference. This does not mean that we have nothing to say about the semantic properties that underlie their language-use. Recall that different specifications of the eligibility-base L gave rise to a variety of candidate reference relations, 'refers$_L$'. We can conjecture that for distant possibilities like the Ectoplasmians, there will be some L or other, such that their language is to be understood in terms of 'reference$_L$'—this wouldn't be reference as we know it, but it would be a *reference-analogue*.[19])

Another possibility is to modify the parochial English-based account. Rather than letting *every* term in English be a reference magnet, we systematically modify it to evade objections: we *filter* English by removing any terms for intentional relations; we *supplement* properties that carve arbitrary distinctions by adding in terms for every other similar property. I think we should also *refine* by taking out terms with false presuppositions, that embody confusions, or which are based on a misconceptions about what does important explanatory work. The first two modifications are designed to avoid the specific objections above of circularity and imperialism. The need for refinement is motivated by the thought that quite generally, theories are improved when stated in the most precise, clear-headed language available. What goes in general goes in particular for the philosophical account of semantic properties. So if we want the best metasemantics, and it's going to be involve heavy use of something like English, the best version will use the best form of the language.[20]

I said earlier that I liked to think of eligibility-based interpretationism as the proposal that selection of semantic theory works by optimizing fit and Humean simplicity; and that is ap-

[18]Perhaps she could be more liberal, and follow the passage from Field quoted at the beginning of this essay, in having the eligibility-makers be the "properties of physics, chemistry, biology, neurophysiology, and those parts of psychology, sociology, and anthropology that can be expressed independently of semantic concepts".

[19]Another strategy here is to see our task in interpretationism as to pick out the realizer of the reference-role. If that's the extent of our ambitions, we leave open that other relations can play the reference-role in other contexts. And if the relevant environments are utterly alien, it's no surprise that we can't specify what that realizer would be.

[20]I suspect that one motivation for what Williamson calls 'the dream of a precise metalanguage' for giving a semantics for vague language is the thought that theories in general are better if we precisify the language in which they're stated. But that assumption seems false to me—sometimes vagueness (in the philosophical sense) may be needed—especially when theorizing about a subject matter which is itself vague. One of the nice aspects of a parochial interpretationism is that it can be nuanced on this point.

propriate when eligibility and Humean simplicity are characterized via the same canonical language. Now, if shoes, string and sealing wax do not have definitions in perfectly natural terms, that's trouble for Humean simplicity (construed as applicable to macro-theories of shoes, string etc), independent of its deployment in metasemantics. I propose we retain the connection, and see our proposals for parochial treatments of eligibility as in addition putting forward corresponding parochial treatments of Humean simplicity.

4.2 Conventionality again

Suppose that the recent versions avoid the specific objections levelled at the crude formulation of parochial eligibility. Are they, for all that, credible as metasemantic theories? It's difficult to feel entirely comfortable with them; they strike us as *ad hoc* or monster-barring. In part, this worry just flags up the fact that the metasemantics is still listiform. Why just physics, chemistry and biology? Why not not include also ecological kinds? Why should English be the starting point for filtering, supplementation and refinement rather than French or Swahili? But we've been through this dialectic. Once the conventionality of the metasemantic theory is acknowledged, we see there's no prospect of a certain kind of 'ultimate' explanation hereabouts. If the theory gets the job done, then the residual question is just whether there's reason to get the job done this way, rather than some other. That's the only explanation of reference-magnetism that's in the offing, but no more is needed.

There is, I think, more extensive 'deep' conventionality in prospect with parochial eligibility than with a naturalness-based metasemantics. I doubt there's anything much to speak for one particular natural language starting point, for example—to take a silly example, whether 'spork' is a reference-magnet will depend on whether we feed in English-at-1900 or English-at-2000. Surely there's no deep metaphysical truth here awaiting discovery.[21] But I don't see this as a problem for the account. Even supposing the various natural-language bases characterize slightly different reference relations, why should that be problematic? We were worried about finding even one reductive account of reference. Having many available would be pleasant.

But one might still hope that not everything goes. In particular, if challenged to explain why the properties of (processed) English rather than their permuted variants count as the reference magnets, it'd be nice to have something to say. Now, it's true that the permuted variant strikes us as perverse, and will be more complex. But sceptics will, with some justice, complain that the decks have been stacked by our parochial understanding of simplicity/complexity, which simply enthroned the properties we happened to be interested in. If we'd started out with the permuted language, we could have analogously reached a permuted criterion of Humean simplicity. What we lack is reason to think this understanding of simplicity is a principled starting point—we lack the Archimedian point that Sider's thesis of the normativity of the natural provided us with, in the original setting.

I don't think having more to say here is a deal-breaker: if the metasemantics proves deeply conventional, even at the level of the choice between English and permuted English, so be it (remember: this doesn't mean that reference is inscrutable, or that there's no fact of the matter what our words mean; it's just that we acknowledge that even keeping usage fixed, concentrating on this notion of reference or meaning is a theoretical convenience rather than anything more deeply rooted). The best chance of avoiding this conclusion, I think, would be

[21] On this view, relative to post-Goodmanian English, grue would be a reference-magnet as much as green. I doubt the occasional gruesome property being magnetic will cause problems, since the lack of Goodmanized predicates for *other* vocabulary in English means that fit with total theory rules out the interpretation of natural language colour predicates as grue. I also don't see much wrong with simply leaving off troublesome artificially introduced predicates from our list—the present approach is by design parochial and unprincipled.

to make the claim that a language that is entirely permuted is just less good than one that talks about things directly. It's better to talk and think directly about shoes, string and sealing-wax than their ϕ-images (in which case, the process of refining and improving the language would, inter alia, involve de-permuting). Perhaps, as Field at times urged, there's something special about intrinsic or non-relation explanations—if so, a language whose every term is extrinsic would in that respect be suboptimal.

Even if deep conventionality was a fact of life, the constraint that we look for the best language in which to formulate the metasemantic theory is not idle. For it tells us pretty unambiguously that we should refine English to get rid of confusions, false presuppositions and the like. This has non-trivial predictions when applied to the original kind of cases that motivated reference-magnetism. Let's go back to the case of 'mass', used confusedly in some community—perhaps ourselves. Now, if we applied the parochial reference magnetism story with English providing the reference-magnets, then all we get is that 'mass' should refer to mass—for whatever the English term picks out, magnetizes. No doubt this is true, but it's desperately non-predictive. The original promise was a view that told us that despite the somewhat confused usage, 'mass' refers to the important explanatory kind in the area, and not to some ad hoc disjunctive property that happens to cleave to usage better. But consider what happens when we take seriously the idea that our metasemantics should be framed in a revised and improved language. Perhaps woolliness over what 'mass' referred to was excusable in our previous state of ignorance, but we now are in a position to see that theories should draw the distinction between rest and relativistic mass, between which our earlier usage equivocated. For the purposes of this just-so story, let's suppose that the frame-invariant notion of rest mass is the most interesting concept, the one that wears the explanatory boots (whether this is the case is a matter of some dispute in literature on physics pedagogy, I gather). We'll take it, in any case, that the best revision of English-as-was is to use "mass" in a way that unambiguously picks out rest mass. Because of this, best metasemantics has it that rest mass is an reference-magnet, and as we'd like, we get the result that our original confused usage in English-as-was did in fact pick out the important kind in the vicinity.

Conclusion

Parochial versions of eligibility-based interpretationism are dialectically important as the maximally non-committal version of the view. If we could do no better, they will satisfice. This permits the advocate of eligibility-based metasemantics to say the following. First, eligibility-based interpretationism, in one form or another, is the right account of the nature of semantic properties. Exactly how eligibility is to be analyzed is an open question, but the form (relative to an input canonical language) is fixed. Since the parochial version is available, and satisfices as a metasemantic theory, there's a bar below which we will not fall. If it turns out that total theory provides richer resources to draw upon (if it provides a suitable set of macronatural properties, for example) then it may be that something like the naturalness-based eligibility is better—but if the world is unkind, we have to take what we get. If naturalness sorts not only properties into the natural and non-natural, but also quantifiers, connectives, objects and the like, then taking the canonical language to be Ontologese (rather than incorporating an essentially arbitrary choice of auxiliary resources) is even better.

If this attitude is right, then the final form of eligibility-based interpretationism is a matter for the philosophical endgame. What turns on the final decision is how *principled* an account of eligibility we can give—and on this depends the extent and type of *conventionality* of reference-magnetism.

References

Barnes, E. J., & Williams, J. Robert G. 2010. 'A theory of metaphysical indeterminacy'. *Oxford Studies in Metaphysics*, **6**.

Chalmers, David. 1996. *The Conscious Mind*. Oxford: Oxford University Press.

Davidson, Donald. 1979. 'The Inscrutability of Reference'. *The Southwestern Journal of Philosophy*, 7–19. Reprinted in Davidson, *Inquiries into Truth and Interpretation* (Oxford University Press, Oxford: 1980) pp.227–242.

Field, Hartry H. 1972. 'Tarski's Theory of Truth'. *Journal of Philosophy*, **69**, 347–375. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 3-29.

Field, Hartry H. 1973. 'Theory change and the indeterminacy of reference'. *Journal of Philosophy*, **70**, 462–81. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 177-198.

Field, Hartry H. 1975. 'Conventionalism and Instrumentalism in Semantics'. *Noûs*, **9**, 375–405.

Field, Hartry H. 1978. 'Mental Representation'. *Erkenntnis*, **13**, 9–61. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 30-67.

Field, Hartry H. 1994. 'Deflationist views of meaning and content'. *Mind*, **103**, 249–85. Reprinted in Field, *Truth and the Absence of Fact* (Oxford University Press, 2001) pp. 332-360.

Field, Hartry H. 2001. *Truth and the Absence of Fact*. Oxford: Oxford University Press.

Heim, Irene, & Kratzer, Angelika. 1998. *Semantics in generative grammar*. Oxford: Blackwell.

Jeffrey, Richard. 1964. 'Review of *Logic, Methodology and the Philosophy of Science*, ed. E. Nagel, P. Suppes and A. Tarski'. *Journal of Philosophy*, **61**, 79–88.

Joyce, James M. 1998. 'A non-pragmatic vindication of probabilism'. *Philosophy of Science*, **65**, 575–603.

Joyce, James M. 2009. 'Accuracy and coherence: prospects for an alethic epistemology of partial belief'. *Pages 263–297 of:* Huber, Franz, & Schmidt-Petri, Christoph (eds), *Degrees of belief*. Springer.

Langton, Rae, & Lewis, David. 1998. 'Defining 'intrinsic''. *Philosophy and Phenomenological Research*, **58**(2), 333–345.

Larson, Richard K., & Ludlow, Peter. 1993. 'Interpreted Logical forms'. *Synthese*, **95**, 305–356.

Lewis, David K. 1969. *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, David K. 1970. 'General Semantics'. *Synthese*, **22**, 18–67. Reprinted with postscript in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 189–229.

Lewis, David K. 1975. 'Language and languages'. *Pages 3–35 of: Minnesota Studies in the Philosophy of Science*, vol. VII. University of Minnesota Press. Reprinted in Lewis, *Philosophical Papers I* (Oxford University Press, 1983) 163-88.

Lewis, David K. 1983. 'New Work for a Theory of Universals'. *Australasian Journal of Philosophy*, **61**, 343–377. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 8–55.

Lewis, David K. 1984. 'Putnam's paradox'. *Australasian Journal of Philosophy*, **62**(3), 221–36. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 56–77.

Lewis, David K. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Lewis, David K. 1992. 'Meaning without use: Reply to Hawthorne'. *Australasian Journal of Philosophy*, **70**, 106–110. Reprinted in Lewis, *Papers on Ethics and Social Philosophy* (Cambridge University Press, 1999) 145–151.

Lewis, David K. 1994. 'Reduction of Mind'. *Pages 412–31 of:* Guttenplan, Samuel (ed), *A Companion to the Philosophy of Mind*. Oxford: Blackwell. Reprinted in Lewis, *Papers on Metaphysics and Epistemology* (Cambridge University Press, 1999) 291–324.

Loewer, B. 2007. 'Laws and natural properties'. *Philosophical Topics*.

Putnam, Hilary. 1981. *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. 1964. 'Ontological Reduction and the world of numbers'. *Journal of Philosophy*, **61**. Reprinted with substantial changes in Quine, *The Ways of Paradox and Other Essays: Revised and enlarged edition* (Harvard University Press, Cambridge, MA and London, 1976) pp.212—220.

Schaffer, Jonathon. 2003. 'Is there a fundamental level?'. *Noûs*, **37**, 498–517.

Schaffer, Jonathon. 2004. 'Two conceptions of sparse properties'. *Pacific Philosophical Quarterly*, **85**, 92–102.

Wallace, J. 1977. 'Only in the context of a sentence do words have any meaning'. *In:* French, P.A., & T.E. Uehling, Jr. (eds), *Midwest Studies in Philosophy 2: Studies in the Philosophy of Language*. Morris: University of Minnesota Press.

Williams, J. Robert G. 2007a. 'Eligibility and inscrutability'. *Philosophical Review*, **116**(3), 361–399.

Williams, J. Robert G. 2007b. 'The possibility of onion worlds'. *Australasian Journal of Philosophy*, **85**.

Williams, J. Robert G. 2008a. 'Multiple actualities and ontically vague identity'. *Philosophical Quarterly*, 134–154.

Williams, J. Robert G. 2008b. 'Working parts'. *In:* Le Poidevin, Robin (ed), *Being: Contemporary developments in metaphysics*. Royal Institute of Philosophy Supplement, vol. 83. Cambridge: Cambridge University Press.

Williams, J. Robert G. 2009. 'The price of inscrutability'. *Nous*, **42**(4), 600–641.