Twentieth Excursus: Reference Magnets and the Grounds of Intentionality

David J. Chalmers

A recently popular idea is that especially natural properties and entites serve as "reference magnets". Expressions refer to these properties at least partly in virtue of their naturalness. This idea sometimes serve as part of a radically externalist approach to reference and content, on which content is fixed by worldly factors that are largely independent of our cognitive apparatus. I think that there is considerable interest in the idea of reference magnetism, but there is little reason to believe in the radically externalist versions of the idea. Still, an analysis of the core issues here can help us to shed light on the grounds of intentionality more generally.

The locus classicus for the idea of a reference magnet is David Lewis's article "Putnam's Paradox". There, Lewis proposes the idea as a response to Hilary Putnam's "model-theoretic" argument against realism. Putnam's argument is itself closely related to Newman's problem, discussed in Chapter 1 and again in the appendix to chapter 7.

As Lewis puts things, Putnam's argument is an argument against a certain theory of reference: *global descriptivism*.[1] Let us say that an expression *E refers wholly by description* when there is an associated description 'the *D*' such that *E* refers to some entity wholly in virtue of that entity's being the referent of a description 'the *D*' that is associated with *E*. A description 'the *D*' is associated with *E* when '*E* is the *D*' is the core of a theory held by users of *E*. For example, 'electron' refers by description if its referent is constrained to be the entity that plays such-and-such role (in physics and chemistry, say) , where 'electrons play such-and-such role' is the core of users' theory of electrons.

[1]Putnam clearly intends his argument to be an argument against an enormous number of theories of reference, not just global descriptivism. But the argument proceeds by arguing in effect that all theory-external constraints on reference can be turned into theory-internal constraints (this is Putnam's "just more theory" point). So one might see the argument as proceeding by argung that any theory of reference can be turned into a version of global descriptivism, and then invoking Newman's argument against the latter.

Here a "theory" need not be especially theoretical: a theory can be any set of sentences that speakers (perhaps "expert" speakers in some cases) are disposed to accept. The notion of a core part of a theory can be understood in various ways. One obvious suggestion is that the core theory associated with E involves only those sentences involving E that are a priori (or perhaps analytic), reflecting those principles that hold over any epistemically possible scenario that users of E might consider. When an expression refers wholly by description, we might say that the only constraints on its referent are *theory-internal*, deriving wholly from speaker's associated beliefs and dispositions to judge.

According to *local descriptivism*, discussed earlier, many or most expressions refer wholly by description, while some other nonlogical expressions, the primitive expressions, do not refer by description but refer in some other way. Global descriptivism says that *all* expressions, or perhaps all nonlogical expressions, refer by description: in effect, the referent of every (nonlogical) expression is determined by some associated theory. Where local descriptivism has both theory-internal constraints on reference (for descriptive expressions) and theory-external constraints (for primitive expressions), global descriptivism says that the only constraints on reference are theory-internal.

Global descriptivism suffers from a by-now familiar problem: Newman's problem. If every nonlogical expression refers by description, then the referent of every expression is determined by an overall theory that can be put using logical expressions alone, and such a theory is almost vacuous. As long as the world has an appropriate cardinality, the theory will be satisfied. Furthermore, it will be satisfied in innumerable ways: any expression can be mapped on to any entity of the appropriate category (so a name can be mapped to any object, a predicate to any property) in a way that satisfies all the constraints of global descriptivism: that is, such that all relevant theoretical claims of the form 'E is the D' come out true.

Lewis draws the conclusion that global descriptivism is false: not all expressions refer wholly by description. It follows that for at least some expressions, one needs nondescriptive constraints on reference. Put differently, one needs theory-external constraints: constraints on an expression's referent that do not derive from users' associated beliefs and dispositions to judge.

At this point, Putnam's argument objects that any further constraints here are "just more theory". In effect, the point is that these constraints are then theory-internal, so that the Newman-style argument applies equally to them. But the mere fact that such constraints can be turned into a theory does not make them theory-internal. The relevant point is that when constraints are theory-external, their reference-determining role in no way derives from speakers' acceptance of

the theory, or from their inclination to make judgments in accordance with the theory. Rather, it derives simply from the truth of the theory.

Lewis considers and rejects one candidate for a theory-external constraint: the causal constraint, holding that an expression's referent should stand in an appropriate causal relation to the expression's uses. Lewis takes the role of causation to be theory-internal. On this view, when the causal constraint plays a role, it does so in virtue of being part of speakers' associated theory. For example, it is part of speakers' tacit theory that Gödel is responsible for our use of 'Gödel': Lewis notes elsewhere that the causal constraint is revealed precisely through speakers' judgments about cases. For other expressions, the constraint is not present at all. Lewis takes all this to recommend causal descriptivism: the thesis that relevant expressions E are associated with a description 'the D' that involves causation, such as (for example) 'the entity causally responsible for my use of 'Gödel''. Given that causation plays a theory-internal role, we have to look elsewhere for a theory-external constraint.

Instead, Lewis endorses another theory-external constraint: the *naturalness* constraint, holding that an expression referent should be relatively natural. Here an entity is perfectly natural when it is fundamental: the perfectly natural entities are the fundamental entities of physics. An entity is more natural than another when it is closer to fundamental: when it can be defined more easily in terms of fundamental entities, perhaps. Then the naturalness constraint says roughly that where there are multiple candidates to be an expressions' referent that are equally well-qualified on other grounds (such as fitting theory), the most natural candidate is the referent. Or considering all expressions at once: the referents of all expressions are determined to be that assignment of entities to expressions that optimizes the combination of naturalness and theoretical fit.

This view has come to be known as *reference magnetism*.[2] The idea is that relatively natural entities serve as "magnets" for expressions to refer to: they are intrinsically more eligible than less natural entities to be referents. When any expression refers to an entity, it does so partly in virtue of relative naturalness: perhaps the relative naturalness of the entity itself, or at least in virtue of the relative naturalness of other entities that serve as referents of related expressions. According to Lewis's *strong reference magnetism*, this naturalness constraint is theory-external and is the only theory-external constraint on reference.

Even if one accepts that naturalness plays a role in reference, there is an obvious alternative to strong reference magnetism: namely *weak reference magnetism*, according to which naturalness

[2] See Hawthorne, Sider, etc.

plays only a theory-internal role. On this view, it is part of our core theory of electrons that electrons are relatively natural, and when naturalness plays a role in determining a theory's reference, it does so in virtue of being part of core theory like this. This view leads naturally to a counterpart of Lewis's causal descriptivism: *naturalness descriptivism*, where naturalness is part of the descriptions associated with relevant expressions. For example, the description associated with 'electron' might be 'The relatively natural entity that plays such-and-such role'.

Interestingly, each of Lewis's reasons for rejecting a theory-external role for causation applies equally to naturalness. On the face of it, the naturalness constraint applies much more in some cases than others. For example, while perhaps 'mass' is constrained to be natural, 'weight' is less constrained (if it were constrained further, it might presumably pick out mass!), while expressions such as 'Canberra' seems less constrained again. This reflects the fact that we are more disposed to regard mass *as* natural. The variation in the role of naturalness and way this variation reflects variation in our attitudes and dispositions already strongly suggests a theory-internal role.

Furthermore, just as the causal constraint is revealed in speakers' judgments about scenarios, so is the naturalness constraint. At least, they are reflected in speakers' informed and reflective judgments: judgments in which speakers are informed about relevant empirical facts (although not semantic facts) and in which they are appropriately rationally reflective. Take a case in which the naturalness constraint seems operative: for example, 'gold' refers to the element *Au* rather than the disjunctive kind involving both *Au* and fool's gold, in part because the element is more natural. Users of the expression were in effect presented with the empirical facts about this scenario and made the judgment for themselves that gold is the simple kind Au. Likewise, the fact that 'plus' refers to the simple operation of addition rather than quaddition is reflected in our reflective judgments about specific pairs of numbers. All this tends to suggest that naturalness is part of speakers' theory associated with these expressions, and that to the extent that it plays a role in determining their referent, it plays a theory-internal role.

So insofar as Lewis's arguments recommend causal descriptivism is over a theory-external causal theory of reference, arguments of the same sort recommend naturalness descriptivism over a theory-external refeence magnetism. This might seem to be merely a *tu quoque* against Lewis—plenty reject causal descriptivism, after all—but I think considerations like this yield a powerful case against strong reference magnetism.

One can put the case as follows. Proponents of reference magnetism (e.g. Sider 20xx) often say that considerations of naturalness can "trump" considerations of theoretical fit: given an entity E with an associated theory T, it can happen that entity a fits T better than b, but that in virtue

of b's greater naturalness, b is the referent of E. If the theory T is constrained to be one in which naturalness plays no role, even weak reference magnetism supports this conclusion. Let us say that *thin theoretical fit* is theoretical fit when a role for naturalness is excluded, whereas *thick theoretical fit* is theoretical fit when any theory-internal role for naturalness is included. Then naturalness can certainly trump thin theoretical fit. But the more relevant question is: can naturalness trump thick theoretical fit?

One can approach this question by asking: in cases where naturalness trumps thin theoretical fit, is the trumping reflected in speakers' (informed and reflective) judgments about the case? Or are speakers' reflective judgments themselves trumped by naturalness? The first option is apparent in the case of 'gold': here speakers' reflective judgments are that the expression refers to the element rather than to the disjunctive kind, so insofar as naturalness plays a trumping role, it is reflected in judgments about cases. The alternative second option requires that trumping is not reflected in judgments about cases, so that speakers reflectively judge that E refers to a less simple entity a, when in fact it refers to b due to b's greater naturalness.

We might call the second option *ultra-strong reference magnetism*: there is not just a theory-external role for naturalism, but one that can trump speakers' judgments about cases. This view is theoretically coherent, but I think that there is little reason to think that it is correct. The main linguistic evidence for reference magnetism comes precisely from judgments about cases: speakers judge that expressions such as 'gold', 'mass', and so on refer to relatively simple entities. If ultra-strong reference magnetism is correct, we should expect that informed and reflective speakers will frequently be simply wrong about the referents of their expressions, and correspondingly about the truth-values of their sentences. To pick an extreme case, due to John Hawthorne (?): it could be that although we take 'Europe' to refer to a highly convoluted area, the expression in fact refers to a rectangular area, so that our judgments about the borders of Europe are largely wrong. I think that this is clearly an unattractive view. If there are reasonable alternatives to this view, we should embrace them.[3]

[3]Some further questions for the ultra-strong view: can simple apparent definitions such as 'vixens are female foxes' can be trumped by naturalness? Can stipulations such as 'glubs are octagonal tables' be trumped by naturalness? If yes, the view is even more bizarre. If no, this suggests that the role of naturalness is not global: this behavior suggests the sort of role one gets on a local descriptivism, where the reference of some expressions is fixed by naturalness and that of others is fixed by descriptions using those expressions.

A related question: are apparently a priori claims such as 'E is the D', where D articulates the theory associated with E that corresponds to the speakers' informed and reflective judgments about cases, truly a priori? Likewise, are scrutability conditionals such as '$PQTI \rightarrow$ Europe is such-and-such' a priori? If reference magnetism is itself a priori,

A nearby view might allow that the truth of reference magnetism will be available to sufficiently informed and reflective speakers, who will then be in a position to make correct judgments about cases: once one knows the correct semantic theory, one will know that Europe is rectangular. If so, naturalness will not trump total theory, but it will at least trump ordinary nonsemantic theory, which is arguably good enough for a theory-external role in a reasonably strong sense. Still, I think a view on which everyone but a few philosophers is badly wrong about the shape of Europe is almost as unattractive as the view in the last paragraph. And once again, the linguistic support for a role for naturalness, just like the support for a role for causation, comes from judgments about cases that are available to ordinary speakers who have never thought about semantic theory. In fact, most of these judgments were originally made by speakers of this sort. So I think this view remains unattractive and unsupported.

Moving to the first option: on this option, cases where naturalness trumps thin theoretical fit are always reflected in speakers' (informed and reflective) judgments about cases, and naturalness never trumps those judgments themselves. On this option, naturalness does not trump thick theoretical fit. Instead, the role of naturalness in determining reference is always reflected in a theory-internal role.

The problem for this option is clear: if naturalness never trumps thick theoretical fit, the obvious conclusion is that naturalness plays only a theory-internal role. The role of naturalness can be encapsulated in naturalness descriptivism: speakers' theories involve claims of the form '*E* is the natural *D*'. Given this much, then at least once the referent of 'natural' is fixed so that it picks out naturalness, any theory-external role for naturalness is rendered redundant.

A strong reference magnet theorist has a couple of options remaining. They might say that although the naturalness constraints are part of speaker's theories associated with expressions, their being part of associated theories plays no role in determining reference. If one has any sympathy for a role for theories or descriptions in determining reference, however, this view seems unattractive. The naturalness constraint meets the standard requirements for being part of a reference-determining theory: in particular, it appears to be reflected in speaker's a priori judgments about cases. Alternatively, the theorist might say that naturalness plays both a theory-internal role and a theory-external role, where the theory-external role is redundant so that reference is overdetermined. This view leaves it unclear why we should posit a theory-external role in the first place,

then presumably some other nearby claims (such as '*E* is the natural *D*') may be a priori instead. Or perhaps reference magnetism is not a priori (just true), and the original putatively a priori claims are only nonconclusively a priori. If so, they are defeasible and may turn out to be false if trumped by naturalness.

though.

Perhaps best, the reference magnet theorist might say that although naturalness never trumps theory, theory leaves certain indeterminacies (highlighted by Newman's problem), yielding ties between candidates that naturalness can settle. Thick theories may determine descriptions such as 'the D' or 'the natural D' associated with every expressions, but naturalness is still needed to settle the overall assignment of referents for the language as a whole.

Still, it is at least suspicious that the external role for naturalness is reflected so neatly in its theory-internal role. The "no trumping, just tie-breaking" view tends to suggest a picture on which naturalness is needed only to settle the reference of a few basic expressions, and then theory can do the rest. In fact, given a view on which naturalness is the only theory-external constraint and on which this constraint is always reflected in theory-internal descriptions, it suffices for the external role for naturalness to settle the referent of precisely one expression: namely, 'natural'. This leads to a picture where naturalness is itself an ultra-natural property, one that is picked out theory-externally by the term 'natural' precisely because it is so natural. Once this is done, the theory-internal role of naturalness can do the rest. This picture is aesthetically pleasing, but it is certainly odd. I think that around this point, it is natural to look elsewhere.

The dilemma here for strong reference magnetism is paralleled by a similar dilemma for a strong causal theory of reference, on which causal constraints play a theory-external role. We can ask once again whether these constraints can trump reflective and informed judgments about cases. If they do, we arrive at an ultra-strong causal theory of reference, which is as unsupported by linguistic evidence and has as bizarre consequences as ultra-strong reference magnetism. If they do not, then the natural conclusion is that causation plays only a theory-internal role, as causal descriptivism holds. The option remains open that it plays parallel theory-internal and theory-external roles, but this option is unattractive as above: it seems that causation need only fix the referent of 'causation', and then theory can do the rest. So I think that just as weak reference magnetism is more plausible than strong reference magnetism, a weak causal theory of reference is more plausible than a strong theory.

Of course Newman's problem tells us that some theory-external constraints on reference are needed. But it makes sense to seek theory-external constraints that are not paralled so directly by theory-internal constraints.

One clue about such constraints comes from the role of theory itself. Even reference magnetism holds that reference is determined by theoretical fit plus naturalness. So the question arises: in virtue of what are expressions associated with descriptions or theories? Here further external

constraints seem to be needed. It is most natural to say that an E is associated with a description D in virtue of its inferential role, or in virtue of speaker's dispositions. But 'inferential role' and 'speaker's dispositions' play very little role in our ordinary theories: "electrons refer to whatever satisfies the inferential role associated with 'electron"' is not folk theory revealed in judgments about electrons in cases. And even if these were part of our theories, the question of how the theories get to be associated with expressions still arise. At the very least, we would need some mechanism to associate 'inferential role' with inferential role, and it is not plausible that naturalness or causation could do this alone: some role for theory would still be needed. I take the moral to be that external constraints are needed to associate expressions with theories, and that inferential role is one candidate for such an external constraint. (This provides another reason to reject Lewis's strong reference magnetism on which naturalness is the only theory-external constraint, and likewise for an analogous strong causal theory of reference.)

Another clue comes from considering a view that Lewis sets aside: local descriptivism. For all Newman's problem suggests, it remains plausible that the reference of very many expressions is fixed by description, and the reference of a few primitive expressions is fixed in some other way. We have seen that models with primitive concepts are attractive in numerous respects, so one might want to look for plausible theory-external constraints that work specifically for these concepts. One may need give some theory-external role to the expression relation, in order to get from primitive expressions to associated primitive concepts (that is, to ground linguistic content in associated mental content). But it is natural to hope that there might be some especially simple story about fixing the reference of primitive concepts.

Even if one rejects definitional models, as I do, the scrutability model suggests a similar moral. On this view, most expressions can be associated with intensions across scenarios, where the scenarios can be specified using a few primitive expressions. If one accepts that these intensions fix reference for the complex expressions, then in effect reference is grounded in (i) whatever fixes the intensions of complex expressions and (ii) whatever fixes the reference of simple expressions. Or if we aim to ground lingustic content in mental content, reference may be grounded in (i) whatever associates expressions with concepts (construed as mental representations), (ii) whatever fixes the intensions of nonprimitive concepts, and (iii) whatever fixes the reference of primitive concepts. There are three corresponding roles for theory-external constraints here. The first role is presumably played by some sort of expression relation (perhaps an appropriate causal/intentional relation), and the second role is plausibly played by something like inferential role. But this leaves open what plays the third role.

What fixes the reference of primitive expressions? We have seen that these are plausibly narrow, and mainly super-rigid, so causal connections to the environment do not seem crucial here. Given the previous discussion, there are two natural initial candidates: (i) acquaintance, and (ii) inferential role. On the face of it, acquaintance seems especially suited for the reference of phenomenal concepts, Edenic concepts of color and spacetime, and indexical expressions such as 'I' and 'now'. On the face of it, inferential role seems especially suited for logical and mathematical expressions, perhaps along with nomic and fundamentality expressions.

Of course 'by acquaintance' is not an especially informative answer to the question of how primitive expressions get their referents. Acquaintance is itself a relation between subjects and referents, and it raises the question: in virtue of what is someone acquainted with something? We might take acquaintance as a primitive relation, in which case this view is not far from primitivism about reference, at least for primitive concepts. Or we might try to explain acquaintance in terms of something more basic, in which case the something more basic will be the more fundamental external constraint.

What might acquaintance be grounded in? Perhaps there might be account in terms of narrow causal role in a cognitive system or inferential role, but I have my doubts, especially given the conceivability of spectrum inversion and the like. I suspect that if a grounding story is to be told, it will give a central role to consciousness. If we are acquainted with qualities such as (Edenic) redness, for example, this is plausibly in virtue of our having certain conscious experiences, including but not limited to perceptual experiences.

This raises the question of the structure of conscious states and what they might be grounded in. If conscious states are grounded in physical or functional states these might thereby also play a role in grounding reference. On my own view, though, conscious states are not grounded this way, and may well be primitive. The question then arises as to whether the primitive structure of conscious sttes is itself relational. On one intentionalist view, perceptual experience fundamentally involves a relation of awareness to certain qualities. If we take such a relation to be fundamental, then again we are not far from primitivism about the most basic referential relations. On another view, intentional structure is itself grounded in nonintentional phenomenal qualities of consciousness. On that view, these qualities may be serving as the ultimate external constraints. I am inclined to prefer the first view to the second, but I view all of these questions as open.

As for inferential role: a detailed story needs to be told about how logical, mathematical, nomic, and fundamentality expressions might acquire their content through inferential role. These inferential roles are likely to be structural inferential roles, so that content is not grounded in in-

ferential links to other specific concepts, but rather in quite general structural patterns of inference that these concepts are involved in. I do not have the details of such an account, but I do not think it is out of the question that such an account can be developed.

Perhaps the biggest challenge to inferential-role accounts is Kripke's plus/quus argument that non-normative dispositions and inferential roles may underdetermine reference. Kripke argues that the same dispositions and inferences might be associated with a badly functioning plus-user and a well-functioning quus-user. Even though the former is disposed to accept '65+78=5', their expression '+' still refers to addition. In response, one might give a grounding role to normative facts about inference, or to what speakers *should* infer: perhaps the difference between the plus and quus user in this case is that the 'plus'-user is being irrational and not judging as they should. Of course normative facts about inference may well be grounded in something else in turn, in which case these grounding facts will help to ground the facts about reference, in a theory-external way.

If we do not want to give normativity a role in explaining intentionality, we might instead give a role to naturalness. One might hold that at least in cases involving nonideal subjects whose dispositions are limited or irrational, reference may be fixed to certain entities (such as addition rather than quaddition) in part because they are so natural. This is compatible with the claim above that naturalness never trumps informed and reflective judgments, at least where reflectiveness involves absence of irrationality, but it gives a role to naturalness in trumping irrational judgments, or in fixing reference in cases where speaker's dispositions are silent. If so, we may have a sort of intermediate reference magnetism, on which naturalness plays a large theory-internal role but also a limited theory-external role, as one of a number of theory-external constraints.

All this suggests a picture on which intentionality is grounded partly in acquaintance and/or consciousness (or whatever grounds them), partly in inferential role, and partly in norms or naturalness. In effect, contents for primitive concepts are fixed by consciousness or acquaintance; contents for nonprimitive concepts are fixed by inferential role plus naturalness or norms (with inferential role fixing dispositions to judge and naturalness or norms extending these dispositions into full intensions). Linguistic content is grounded in this intentional content along with the expression relation.

These factors might be incorporated into theories, but in many cases they are not part of speakers' associated theories (unlike the roles for causation and naturalness considered earlier), and in any case they do not play their roles in virtue of being part of these theories. All of these factors will play a theory-external role, but none of them are radically external, in that their role will never trump informed and (ideally) reflective judgments about reference.

An enormous amount of work is required to turn a brief sketch such as this into a full picture of the grounds of intentionality and of linguistic content. But the framework developed here might at least illuminate what needs to be done.