How to Be a Meaning Holist<*>

Eric Lormand
Univ. of Michigan
email: lormand@umich.edu

The Journal of Philosophy, January 1996

Meaning holists hold, roughly, that each representation in a linguistic or mental system depends semantically on every other representation in the system. The main difficulty for holism is the threat it poses to meaning stability--shared meaning between representations in two systems. If meanings are holistically dependent, then semantic differences anywhere seem to balloon into semantic differences everywhere. My positive aim is to show how holism, even at its most extreme, can accommodate and also increase meaning stability. My negative aim is to provide reasons for rejecting various nonholist proposals, at least for systems of mental representations.

1 What is meaning holism?

Meaning holism is a theory about the interrelatedness of meanings within a representational system.<1> On meaning holism about a system *S*, virtually every representation's meaning in *S* depends on that of virtually every other representation in *S*. By contrast, a "nonholist" claims that a representation does not depend semantically on virtually all others in *S*, and a "localist" claims more particularly that a representation depends semantically on virtually no others in *S*.

Setting aside for now the 'virtually' hedges, arguments for meaning holism tend to have two phases, an "all-or-none" phase and an "at-least-some" phase. Let *r* be a representation in a system *S* containing other representations inferentially related to *r*. The all-or-none phase argues that there is no (sharp or vague) distinction between (i) the other representations in *S* relevant to or constitutive of *r*'s meaning and (ii) those irrelevant to or merely collateral for *r*'s meaning. In short, there is no constitutive/collateral distinction. The suggested consequence is that *r*'s meaning depends either on all the representations in *S* or on none of them. Reasons to reject a constitutive/collateral distinction are reasons to accept either extreme holism or extreme localism. The at-least-some phase, next, argues against the *complete* independence of *r*'s meaning from its inferentially related representations. Reasons to reject localism become reasons to accept holism.

What reasonable constraints on theories of meaning might underwrite such an argument? I think we can find some by focusing on a notion of meaning that is suitable for use in cognitive sciences such as cognitive psychology, cognitive neuroscience, psycholinguistics, and artificial intelligence, as they might (or might not) be opposed to common sense. By and large, cognitive scientists are concerned primarily with mental representations (e.g., contentful psychological items, perhaps akin to beliefs and ideas), and only secondarily with public-linguistic representations. Against eliminativism, behaviorism, and instrumentalism, I will assume without argument that there really are mental

representations (though perhaps we haven't discovered all the types, and perhaps the types we currently posit don't exist), and that they are causal intermediaries between perceptual and motor organs. Nevertheless, I will try to avoid controversial claims about their specific form or function.<2>

I take it that, at a minimum, cognitive science appeals to the meaning of mental representations for three theoretical purposes: (i) to help specify the mental representations, (ii) to help express generalizations about their (actual or idealized) functional role--generalizations relating mental representations not only to external conditions, via processes such as perception and action, but also to other mental representations, via processes such as inference, and (iii) to help explain these generalizations. I will be interested primarily in the pursuit of these goals--in *psycho*semantics rather than *linguo*semantics or semantics as a whole.<3> In order to fulfill the purposes of psychological explanation, psychosemantic kinds should be psychological natural kinds rather than kinds arbitrary with regards to (actual or idealized) psychological generalizations. What I will argue, on the basis of some plausible and live psychological speculations, is that for a typical mental representation *r*, there is no (sharp or vague) psychological natural kind--relevant to (i)-(iii)--that should be used to distinguish some representations relevant to *r*'s meaning from some representations irrelevant to *r*'s meaning. The phenomena best explained by *r*'s meaning are best explained either by *r*'s relations to all other representations in *r*'s system or by *r*'s relations to none of them. Hence meaning, in whatever sense best serves cognitive-scientific explanation, is either holistic or localistic.

To reach holism, this all-or-none phase of the argument must be supplemented by an at-least-some phase, that is, by support for what has become known as 'inferential role semantics' (IRS).<4> IRS holds that the meaning of a mental representation depends at least partially on other mental representations that are inferentially--e.g., causally, functionally, and rationally--related to it.<5> A full IRS theory of meaning may appeal to more than inferential relations, such as causal relations between mental representations and nonmental phenomena. The main alternatives to IRS are purely "referential" (or "denotational") theories of meaning, according to which such mind-world relations exhaust meaning.<6> Purely referential theories claim that meaning is *fully determined* by referential relations, so that any two representations have the same meaning if they corefer--if they are about the same existing objects, properties, or facts--or if they have the same "reference conditions"--if they would corefer however the world might be.

Perhaps the main reason for preferring IRS to purely referential theories is that IRS underwrites psychological purposes for being interested in meaning, by tying meaning partially to the psychological role of mental representations. Many philosophers resist referential theories on the grounds that meanings should be individuated more finely than referents or reference conditions, if meaning is to be of service to psychological explanation. Psychologists appeal to meanings to specify groups of mental representations that (allegedly) enter into interesting psychological generalizations. However, referents and reference conditions do not appear sufficient for this purpose. On plausible assumptions, for example, [gold] and [Au] have the same referent, since gold is identical to (the element) Au. They even have the same reference *conditions*, since no matter what way the world is or can be, either both refer (to the same thing) or neither refers. Nevertheless, although there might be some distinctive psychological generalizations shared by virtually all people with a [there's gold in them thar hills]-belief, there might be few if any shared by virtually all people with *either* this belief *or* a [there's Au in them thar hills]-belief, given that so many people have not come to accept [gold = Au]. Given that referents and reference conditions alone are not individuated finely enough to reflect

these differences among mental representations, inferential role semantics steps in to provide the apparently missing ingredient in meaning--[gold] and [Au] differ in meaning by differing suitably in inferential role. Alternatively, a purely referential theory can insist that [gold] and [Au] do have the same meaning, and hold that inferential role *supplements* rather than *shapes* meaning, in drawing psychologically interesting distinctions.<7> While the resulting situation has many earmarks of a philosophical standoff, there is some room for argument. The most pressing issue is whether an IRS can avoid meaning holism, or at least the damaging objections to meaning holism.

Before pursuing this issue, my description of holism requires sharpening along at least two edges: (a) what *kind* of dependence among meanings the holist posits, and (b) *which* meanings the holist claims to be dependent. To make my defense of meaning holism as difficult as possible, I will construe it very boldly--indeed, more boldly than any version that I have seen described, much less defended. As to the kind of dependence, I will suppose that the holist requires "differential" dependence rather than mere "influential" dependence. '*X* differentially depends on *Y*' is stronger than '*X* is influenced by *Y*' in requiring that, necessarily, if *Y* were different, *X* would be different.<8> For example, if M('cow') and M('brown') are differentially dependent meanings expressed in English, then were English somehow to cease having representations with M('cow'), it would necessarily cease having representations with M('brown'). (More crudely, if 'cow' changes meaning, then 'brown' must, too.) As to the *extent* of dependence, I will focus for the time being on the extreme holist view that the meanings of *all* the representations in a system are interdependent. Although I will defend even this extreme view against standard objections, slightly less extreme positions should count as vaguely holistic. But the extreme view is simpler for my present purposes of describing the troublesome ramifications of holism.

2 Troubles for holism

The most serious objection to holism is based on stability of meaning across representational or inferential differences. If meaning holism about a system *S* is true, then a change in meaning of any representation in *S* requires a change in meaning of *all* representations in *S*. If 'carburetor' in English changes meaning, so do 'cow' and 'brown,' since their meanings differentially depend on it. If one's [carburetor] idea changes meaning, so do one's [cow] and [brown] ideas. Furthermore, any gain or loss of a belief involving an idea (e.g., a belief [aunt polly owns a cow]) changes the meaning of that idea (e.g., [cow]). Thus meaning holism threatens intrapersonal meaning stability, for it seems that no mind or ideolect at different times could have representations with the same meaning, given that no one has *all* the same ideas or beliefs (or, for idiolects, speech dispositions) at two times. This would mean that intrapersonal, diachronic *agreement* never occurs. It would also mean that such *dis*agreement never occurs, since for a mind to believe *that p* at one time and to believe *that not-p* at another time, presumably it must have the ability to represent *that p* at both times. To take the extreme case, it would be impossible to perform even simple deductions without equivocation, since the meaning of one's terms would change in the course of formulating and accepting new premises or conclusions. The situation doesn't look much better in the interpersonal or cross-linguistic case. Meaning holism seems committed to the claim that if two languages or minds differ semantically at all, they differ everywhere, in the sense that they cannot have any representations that share meaning. This would threaten both interpersonal agreement and interpersonal disagreement, on the plausible assumption that no two people have all their beliefs (or other mental representations) in common, and no two languages have precisely the same overall expressive powers.

So holism threatens to lead to meaning instability, and (according to the all-or-none considerations to be developed in sections 3 and 4) inferential role semantics threatens to lead to holism. This provides the main reason for resisting IRS in favor of purely referential theories (see section 1). Referential theories are attractive because they promise to provide a large amount of meaning stability. For example, two people who are exposed to the same object or kind may have ideas about it--and so, on a referential theory, ideas with the same meaning--even if they disagree *completely* about it. At a minimum, sameness of reference and meaning may withstand *some* differences in associated beliefs. So a purely referential theory is one way to avoid the problems for meaning holism. If there is no way for IRS to delimit (even roughly) the associated representations that contribute to the meaning of a representation, IRS seems forced to meaning holism, with extreme instability. This would be a strong reason to prefer purely referential theories of meaning, if we cannot provide for meaning stability within a holistic version of IRS. But if a holistic IRS can be *compatible* with stability, the main objection to IRS is blocked. Furthermore, if as I believe a holistic IRS can provide *more* stability than a referential theory, the main consideration against IRS becomes a positive argument in its favor.

Here is the strategy I recommend to the holist for securing meaning stability. The root worry is that the holist commitment (1) entails the instability claim (2):

(1) if a representation in a system *S* changes meaning, all representations in *S* change meaning.

> (2) if a representation in *S* changes meaning, no representation in *S* has the same meaning it had before the change.

But the inference from (1) to (2) is based on a certain implicit presupposition about representations and meanings. Essentially, the presupposition is that there is a one-to-one correspondence between token representations and meanings, so that token representations are unambiguous meaning-bearers. The apparent inconsistency between holism and stability depends on (2)'s talk of "the" meaning of a representation, construed as the *single* meaning. But if a representation has *multiple* meanings, all at once, the inference from (1) to (2) is blocked. (We can construe 'the meaning' of a representation as numerically indefinite, covering all a representation's meanings-with-an-*s*, just as we speak of 'the effect' of an event meaning an indefinite number of its effects-with-an-*s*.) When a representation (e.g., 'carburetor') changes meaning, each other representation (e.g., 'cow' or 'brown') can change some of its meanings--satisfying (1)--without changing all of its meanings. Each representation still can have some of the same meanings as before--falsifying (2). This reconciles meaning holism with meaning stability.

To illustrate this, I need to introduce more details about multiple meanings, although it is not my purpose in this paper to provide a full development and defense of the multiple meanings view (MM).<9> Suppose that for any given (token or type) representation *r* in *S*, the set of other representations in *S* (with which *r* is potentially used in inference) can be divided into several (suitably characterized, perhaps overlapping) "units" such that *r* expresses several meanings, perhaps one for each of these units. For purposes of illustration, think of each unit for a representation as a separable rough test for the acceptable use of that representation.<10> To take a simplified example, suppose that a small child, Larry, uses a token mental representation [bird] with the following three tests among many others: [feathered flying animal], [thing similar enough to B1,...,B*n*] (where the

[B*i*] represent alleged birds), and [thing called ‘a bird’ by mommy]. On the view I propose, each unit helps to yield a meaning--as opposed to "the" meaning--of [bird].<11> Suppose Larry forms a belief involving (acceptance of) the mental representation [this is a bird]. On the multiple meanings view, acceptance of a mental representation realizes multiple beliefs simultaneously; in this case, perhaps, when Larry accepts [this is a bird] he thereby believes *that this is a feathered flying animal*, *that this is a thing similar enough to B1,...,B*n, *that this is a thing called ‘a bird’ by mommy*, etc. His *degree of belief* in each of these propositions depends, presumably, on how strongly he accepts [this is a bird], and on how strongly he associates [bird] with the three tests. We would ordinarily, and independently, take Larry to have each of these beliefs, at least as "implicit" (or, in one sense, "tacit") beliefs. This suggests that MM, perhaps surprisingly, does not have implausible consequences for attitude ascription.

MM can be accepted without accepting meaning holism; even if *r* has multiple units, taken together these may not contain every other representation in *S*. But if each other representation does find its way into some unit(s) for *r*, like proper holists we don’t have to distinguish between representations that are and aren’t meaning-relevant to *r*.<12> Each unit would provide a separate meaning for *r*, and would be vitally relevant to that meaning, and completely irrelevant to others. But the *total* set of meanings of *r*--i.e., its "meaning" in the numerically indefinite sense--would depend (differentially) on each unit. Via some meaning or other, *r* would depend semantically on every other representation in *S*, which would satisfy meaning holism for *r*.

Nevertheless, we would also be able to secure a large degree of meaning stability. Two representations may have different *total* sets of units but *still* share meaning, if the sets overlap, or share at least one unit. Even after a change in one unit, a representation still has many of the same meanings it did before. Also, two people in this situation can literally share beliefs, and can literally disagree. To continue the simplified example, suppose that Larry’s younger sibling Moe also uses a mental representation [bird] with the test [thing called ‘a bird’ by mommy], although without tests based on feathers, flying, and sample birds. (Perhaps Moe is blind and uses [animal that sings like S1,...,S*m*], where the [S*i*] represent alleged bird songs.) On MM, the two children can genuinely agree and disagree about birds, despite the holism. For example, when Larry applies [bird] to an object but Moe withholds [bird], they may genuinely be disagreeing about whether the object is a thing called ‘a bird’ by their mommy.

The stability of this specific meaning depends on the fact that the two children both appeal to mommy; if a third child Curly is like Moe but uses [daddy] instead of [mommy], then for all I have said Curly’s [bird] and Larry’s [bird] do not share any meanings. Against this, plausibly, two people who interact with the same object or kind can have representations about it with the same meaning, even if they share *no* tests for the representations in question. As I mentioned above, this would be an attractive consequence of any purely referential theory according to which coreference entails synonymy. But this additional stability is easily accommodated on MM. Given that a representation has a long list of meanings anyway (i.e., those yielded by its units), we can simply *add* referents to this list of meanings. In this way, some of Curly’s meanings--say, the *set* of things called ‘a bird’ by daddy, or the *property* daddy picks out--can be identical to some of Larry’s meanings--say, the set of things called ‘a bird’ by mommy, or the property mommy picks out--even though their words have completely different tests. Such a theory would *at least* provide all the meaning stability of a referential theory, since every meaning posited by the referential theory would also be posited by the

holistic IRS.

Furthermore, an IRS with multiple meanings would *also* provide all the psychologically-relevant meanings that purely referential theories omit (see section 1). In this way, IRS can provide *more* meaning stability than a purely referential theory. Both kinds of theories can describe two representations (in different minds, or in one mind over time) as precisely sharing referential or "coarse-grained" meaning. But when sets of tests *do* partially overlap, the multiple meaning view reflects the resulting psychological similarities by allowing two representations (with different but overlapping inferential roles) precisely to share "fine-grained" meanings (a.k.a. definitions, Fregean senses, or modes of presentation). Purely referential theories have no truck with fine-grained meanings. For this reason, the multiple meanings view can be more *precise* than referential theories in comparing two minds (or one mind over time) for semantic identities and differences.<13> We seem forced to choose between having our cake (stable meaning) and eating it (sacrificing it for psychological precision). Why make this choice, when we can have many cakes and eat many others?

3 Concepts and stereotypes

Of course, a reconciliation of holism with stability is of more interest if there is some reason to *accept* stability and holism. I think meaning stability is self-evidently a good thing if we can get it–though some IRS theorists despair of securing literal stability,<14> no one touts instability itself as a desideratum. By contrast, I don't know any general reason why we should *want* meaning holism to be true, although in the rest of this paper I will try to describe an important theoretical standpoint from which it, or an approximation of it, is plausible and useful. That standpoint begins with inferential role semantics, cast as a theory of meaning for cognitive science (see section 1). The most important question for an inferential role semantics is: which inferential relations matter to meaning? Of course, if one insists on reaching a nonholistic IRS, there are innumerable ways to restrict the class of meaning-relevant inferential relations. Any arbitrary method of ordering the inferential relations of a representation, coupled with any arbitrary method of dividing the series in two, could be used to generate a meaning-relevant vs. meaning-irrelevant distinction. But I am assuming that semantic kinds and distinctions are nonarbitrary from the point of view of cognitive-scientific explanation, and that they at least in this respect track psychologically interesting natural kinds and distinctions. If we accept a restriction that is not psychologically natural, we risk compromising the potential role of meaning in psychological explanation, and so we compromise the main reason to accept inferential role semantics rather than purely referential semantics. If there is no psychologically natural restriction, then, I think inferential role semanticists ought to try to live with holism. The considerations to come hopefully make plain why an IRS cannot dismiss meaning holism lightly, despite its problems.

I will restrict attention to those mental representations commonly called 'concepts' in psychology.<15> Psychologists tend to use this term for representations that perform two main functions: (1) a concept serves as an "address" in memory at which various types of information (i.e., other representations) may be organized, stored, and retrieved, and (2) a concept serves as some sort of constituent or determinant of semantically "complete" mental representations, and thereby helps to determine the inferential role of such representations. Often, this usage is restricted in two ways: (3) a concept serves as a mental *predicate* or *common name* rather than, say, a proper name or a logical particle, and so is used for *categorization*, and (4) a concept is a part of one's "higher" mental

processes rather than one's "lower" processes such as early vision and late motor control. A psychologically motivated theory of meaning would hopefully apply to all mental representations, but it is reasonable to begin with psychological speculations about these central cases. Specifically, I will seek to describe (actual or idealized) inferential relations that establish (good or bad) *categorization procedures* for a concept--representational tests people use to help apply the concept to some specified thing.

The most natural way for an inferential role semanticist to avoid meaning holism is to appeal to so-called "classical" categorization procedures, restricted tests that might be supposed to provide necessary and sufficient conditions for the applicability of a representation. Specific examples of classical definitions tend to be very controversial, but a short list can at least illustrate what nonholists have traditionally *hoped* for. A system might test for the applicability of [aunt] using [(sister of a parent) or (wife of a brother of a parent)], for [bicycle] using [2-wheeled pedaling land vehicle], for [cat] using [animal that is or was a kitten], or for [gold] using [element with atomic number 79]. If enough such tests form a nonarbitrary psychological kind, this kind might be used to generate a nonholist theory of meaning. Much of my discussion below will turn on searching for something distinctively interesting about such categorization procedures, from the point of view of psychological explanation.

Psychologists typically search for the procedures most important for normal, everyday categorization--presumably these yield the bulk of the generalizations about categorization. In order to account for these generalizations, then, psychologists need to appeal to natural features of inferential relations such as *salience*, *availability*, *recent usefulness*, or sheer *frequency* of use. But *these* explanatory properties are not distinctive of classical definitions. The firm consensus from psychological research into categorization is that people normally use rough-and-ready *stereotype-based* procedures rather than classical definitions.<16> These procedures use representations of several forms. They can use lists of properties that one thinks the members of a category at best *tend* to have, lists that one uses as mere *heuristics* for categorization: [bonneted kisser of an uncle] might be a stereotype for [aunt], [thing in a box marked 'schwinn'] might be one for [bicycle], [purring mouse chaser] for [cat], or [precious yellow metal] for [gold]. One can also categorize by determining whether something has a sufficient (weighted) amount of the properties in a certain list [F1,...,F*n*], or, as it might be, whether something meets the description [thing with enough of: F1,...,F*n*]. Finally, one can categorize by comparing something with specific stereotypical *exemplars* of a category, rather than listing alleged properties of exemplars. Suppose that someone represents a few exemplary chairs [chair1,...,chair*n*]. In this case, he may apply [chair] to something if it meets the description [thing similar to chair1,...,chair*n*].<17> A single category can be associated with several such stereotype-based procedures, as well as nonstereotypical procedures.

Classical definitions tend to score low on such psychologically important measures as frequency of use and salience in categorization. Nevertheless, someone especially interested in the psychological importance of stereotypes might offer these measures as a criterion for determining which representations affect meaning. On this view, a representation would mean the same as the representation involved in its most salient or frequently used categorization test. This is implausible, however. Suppose we *explicitly stipulate* that a new word 'superlune' is *defined* as 'heavenly body more massive than the moon.' Since it is very hard to "weigh" heavenly bodies, we might decide that size is a reasonable, though fallible, indicator of mass. So suppose we begin using [larger than the

moon] as what we take to be a quick-and-dirty test for superlunehood, and as it happens we apply this test more frequently than we do measurements of mass. Naturally, there would be some cases--perhaps big gas clouds--in which we would override the size test. However, on the maximum-frequency or maximum-salience proposal, [larger than the moon] would have become constitutive of [superlune]'s meaning. On this view, [more massive than the moon] would be irrelevant to the meaning of [superlune], and so presumably we would be *mistaken* when we occasionally override size, even though overriding is consistent with our stipulation and we don't have the slightest inclination to consider overriding as mistaken.

Intuitively, the mass test matters to the meaning of [superlune] because it is in some sense psychologically *stronger* than the size test, even though the size test is more salient and applied more frequently. Much of the hope for a (psychologically motivated) nonholist inferential role semantics rests on explaining what kinds of strength might be relevant to drawing a meaning-relevant vs. meaning-irrelevant distinction. While there is no way to canvas all the possible construals of strength, here are two kinds of psychological phenomena that might be taken to constitute or at least reflect the degree of strength of a test: *commitment strength*, or degree of resistance to abandoning the test, and *imaginative strength*, or degree of resistance to imagining how the test could go wrong. Even though the size test for [superlune] is normally used instead of the mass test, it might be weaker: we might more easily abandon the size test, or more easily imagine in detail how it could go wrong. Since strength admits of degrees, in order to use it to draw a constitutive/collateral distinction we would need a psychologically nonarbitrary way of dividing the dimension into two (sharp or vague) parts. For example, perhaps only the *maximally* strong tests matter to meaning, or perhaps all the *comparatively* strongest do.<18>

We thus have several proposals to consider about the psychological relevance of semantic kinds and distinctions. How can we test them? I will focus on *abnormal*, imaginary categorization tasks rather than the psychologist's experiments designed to illuminate normal categorization. In order to measure the relative strength of the various tests associated with a representation, it is useful to imagine the tests yielding clearly different verdicts, which they don't normally do. This is one of the primary roles of thought experiments in philosophy of meaning, a role which may qualify them as preliminary, questionnaire psychological experiments. It remains to be seen whether genuine psychological science can be made of these initial forays.<19> Nevertheless, I will be extracting some psychological speculations about strength from the philosophical literature on meaning. This is because I think that this literature contains interesting supplements to the psychological literature on our thoughts about categories and our procedures for categorization. The psychological claims I will make are not beyond dispute, but I think when we see the details the claims will seem coherent with and initially as plausible as the claims available from the psychological literature on categorization.

Even if the use of bizarre thought experiments is potentially respectable as psychology, since I will be arguing from counterexamples and not from general principles, I need to block several suspicions about the import of my arguments. First, it would be a mistake for the holist to rest on *borderline* examples, with an eye toward establishing the vagueness of the boundary between a representation's constitutive and collateral tests, or the vagueness of the boundary between representations with holist and nonholist meanings. Vagueness is not a problem for the nonholist. If the set of *clearly* constitutive tests for a representation is small enough, or if the set of clearly nonholist representations is large enough, holism is false regardless of the borderline cases. My examples need to bear on apparently

clear cases of representations with nonholist meanings, and relatively clear cases of constitutive or collateral tests for them. Second, it would be a mistake for the holist to rest on examples of psychological *breakdowns* or other limitations, with an eye toward arguing that no tests are maximally strong or comparatively strongest. We can of course abandon a test by carelessness, or confusedly imagine it going wrong while on drugs, etc. The nonholist may appeal to kinds of strength which reflect substantial psychological idealizations away from actual, normal, or botched categorization practice.<20> I don't suppose there is anything psychosemantically illegitimate about this; the role of semantic kinds and distinctions may be to describe and explain highly idealized generalizations about the causal role of representations. My examples need to be free of psychological breakdowns, regrettable failures of rationality, limitations on some independently important resource such as memory, attention, or reflection, or mere performance slips. If they are free of these factors, the holist will need serious explanatory reasons for idealizing away from my examples.

Let's begin with reasons against believing that people typically treat tests as maximally strong, in the commitment sense. Call this case, due to Elliott Sober, *Wise One*:

> Suppose we know someone who is an extremely insightful and trustworthy authority. We also know this person to be very honest. This Wise One says to us one day, "philosophers are always saying that all bachelors are unmarried. But if you look very carefully at what these concepts mean, you'll see that this doesn't have to be true. And, in fact, there are some bachelors who are not unmarried." . . . [I]t would be *pigheaded* to simply dismiss the remarks of the Wise One out of hand. People have made mistakes in analyzing concepts before, and if the Wise One is so smart and honest, we ought to take him seriously in the present case. So we may decide to back off from our belief.<21>

I assume that [bachelor] *seems* clearly a representation with a nonholist meaning, and that its inferential relations with [unmarried] *seem* clearly constitutive of its meaning. The example is not unfairly borderline. But as a matter of psychological speculation, I expect that most people of sound mind and body, in the envisioned circumstances, *would* back off even from the belief that all bachelors are unmarried. People tend to be modest about their use of words and ideas, in the face of acknowledged experts. And I do not think this influence should be idealized away as a breakdown, limitation, arbitrary interference, or slip. I agree with Sober that even with care and reflection, as a rational inference from the evidence, one *should* back off from the belief. If someone would refuse to back off from the belief, it would be more tempting to try to idealize away from factors such as dogmatic inertia or inability to trust others.

One drawback of Sober's example for my purposes, however, is that the envisioned change in the use of [bachelor] is quite minor. One would presumably still use unmarriedness and maleness in tests for bachelorhood, and simply be prepared to add a few odd further tests to handle a few odd special cases. Many people already allow for this in suspecting that the Pope is not a bachelor, or in suspecting that married playboys in polygamous cultures are bachelors. It would be much more telling to imagine the Wise One claiming that there are bachelors, but *none* of them are unmarried men! I think Sober's style of case would work even for this belief, but it would be considerably harder to push. Even so, these sorts of thought experiments do not bear on the *imaginative* strength criterion of meaning-relevance.

We may be able barely, given a modified *Wise One* case, to imagine *that* no bachelors are unmarried men, but can we imagine *how*, in detail? If not, this limitation on imaginability would be a genuine psychological phenomenon worth explaining, and a natural (beginning of an) explanation would be that [bachelor] means roughly the same thing that [unmarried man] means.

Against this, I think that we *can* imagine how it can be that no bachelors are unmarried men. We are not limited to appeals to the Wise One's authority. Here is a start; call it *Mars*:

> Suppose that unbeknownst to us all the unmarried people on Earth are female, though roughly half of them are very cleverly disguised (even from themselves) as males, and we have been calling them 'bachelors' all along. Furthermore, just as these alleged bachelors are about to say 'I do,' they are instantly and secretly killed by Martians and replaced by human males. (Perhaps these newly married males have been raised on Mars from humans abducted in prehistoric times.)

I think this is a case in which we are prepared (rightly or wrongly) to discover *empirically* that bachelors are not unmarried males at all, but unmarried, disguised females earmarked for replacement. Similarly, we can imagine how all bachelors could be married. Call this case *Venus*:

> Suppose that each male infant on Earth is in a secret flash carried off and wed to the polyandrous Queen of Venus, and re-released on Earth to vie for the coveted Earth title of 'bachelor.' Suppose also that there are long-forgotten legal clauses, coincidentally in all Earthly constitutions, giving some official recognition to such transplanetary marriages.

Perhaps in this case there is some inclination to hold that there are no bachelors, and also some inclination to withhold judgment. But on reflection, I think that we would and should take ourselves to have discovered, empirically, that our (good old) bachelors are and have always been married. We would still have every reason to keep track of differences between these men and those with Earth-wives, and [bachelor] would be as effective as ever in doing so. After all, we are accustomed to discovering that we are wrong about the "hidden" properties of entities we categorize together. We do have *very* strong beliefs that all (many, some) bachelors are unmarried and all are men. But these beliefs are not *maximally* strong either in the commitment sense or in the imaginative sense. They do not have *these* psychological properties to distinguish them from our very strong but apparently meaning-irrelevant beliefs that all (many, some) bachelors fear commitment and all are smaller than the moon.

As always, there is an understandable tendency to discount such considerations as "merely psychological" and *obviously* irrelevant to semantics. There may or may not be semantic concerns that are best shielded from psychological concerns, but *psychosemantic* concerns clearly should not be shielded. Still, perhaps a good psychological theory of our *idealized* inferential competence will uncover nonholistic restrictions on imagination and belief change, restrictions which are sometimes relaxed in practice due to such interferences as fatigue, bewilderment in the face of bizarre thought experiments, forgetfulness, etc. If so, perhaps the thought experiments I have described are irrelevant even to psychosemantics.

It is hard to assess this proposal in advance of evidence about the details of idealized psychological competence. At any rate, I want to suggest that our reactions to these thought experiments--especially our apparent modesty and flexibility about even very strong beliefs--*may well* turn out to be a deep feature of our inferential competence, even idealizing away from breakdowns such as fatigue, bewilderment, and memory limitations. Modesty is part of a *rational* strategy for fixing beliefs and decisions, since, ideally, it enables inferences to take place *all things considered*. From this standpoint, it is *restrictions* on inferential dispositions that are psychological breakdowns, born of lack of memory, lack of time, lack of imagination, the dogmatism of the familiar, etc. (I have in mind our tendency to *insist* as follows: 'Bachelor' means 'unmarried man,' period! End of story! Leave us alone!) So my claim is not that we can be *fooled* into breaking the connections between [bachelor] and [unmarried man]. It is that, contrary to initial appearances, there are perfectly possible ways our world could be such that given full access to all the evidence the world has to offer, and making no independent false assumptions, the rational conclusion upon full reflection, free from semantic restipulation, is that none of the bachelors are unmarried men.

Can this claim be extended indefinitely to other apparently clear cases of meaning-relevant inferential connections, or are the [bachelor] examples somehow uninstructive? I think we can generalize, because the thought experiments turn on very widespread features of categorization practices. Thought experiments such as *Wise One* turn on a practice of deference to authority, which is generally present at least for concepts that are associated with public-linguistic expressions. The *Mars* and *Venus* thought experiments turn on a practice of using alleged samples and generalizing from their potentially hidden properties, a practice which is present at least for concepts of natural kinds, but which I believe is much more widespread. In the next section, then, I want to bring these practices into the forefront in our search for categorization tests that score high in psychological strength.

4 Experts and natural kinds

A child learning the word 'cat' might associate with [cat] the representation [thing called 'a cat' by mommy]. Similarly, if in conversation I "drop" a new name 'McSmith,' you may form a new representation associated with [person called 'McSmith' by Eric]. Even after coming to a rich set of beliefs about something, we still associate such metalinguistic representations with it, typically with loose appeal to other language users: [. . . doctors], [. . . most French speakers], [. . . enough people I talk to], etc. Although these metalinguistic representations are not much studied by concept psychologists, they are very frequently used in categorization, e.g., anytime one categorizes something on the basis of someone else's linguistic testimony.

What may be more surprising is the psychologically strength these categorization tests have in comparison with stereotypes or even the stock examples of classical definitions. (In this section, I will use 'strength' indifferently for resistance to rejection and resistance to imagined inadequacy.) This is suggested by typical reactions to thought experiments in the philosophical literature. If you come to think that you're the only one around who applies the word 'cat' to ferrets (in addition to lions, tigers, and housecats), or that expert zoologists have been hiding the dirty secret that what they call 'cats' *don't* purr (but instead have a cute way of exhibiting their flatulence), you're likely to alter your [cat] stereotypes *rather* than the metalinguistic representation. Sober's *Wise One* thought experiment illustrates the same phenomenon for 'bachelor,' and variants could be constructed for virtually any

concept with metalinguistic tests. This is an example of Hilary Putnam's "division of linguistic labor," or the strong tendency we have to defer to our linguistic community (particularly acknowledged experts) in categorization.<22> This tendency is not only *actual* but plausibly *rational*, and not to be idealized away as due to interferences on a fundamentally noncooperative competence.

Normally, we think that even the experts can be radically, and forever, mistaken about how to categorize things. This might be because we strongly associate [cat] with a representation like [thing of a natural kind that best fits enough alleged-cats]. So, we might discover that zoologists are mistaken, and the familiar things [cat] has been applied to aren't animals after all, but instead robots.<23> Many people react to such thought experiments by continuing to apply [cat] to these robots, apparently favoring the natural kind test over tests involving the concept [animal]. Although these claims are most plausible as claims about actual psychological dispositions, they seem to survive as claims about idealized dispositions, abstracting away from limitations on reflectiveness, idiosyncrasies, etc.

I think that the *Mars* and *Venus* thought experiments on [bachelor] also turn on our dispositions to generalize from recognized samples, which are essentially the same dispositions involved in natural-kind categorization tests. How can such dispositions bear on concepts apparently *not* of natural kinds? First, it is important to keep in mind that the term 'natural kind' is a technical philosophical term that few people use, and that there are several open and interesting questions about how people think of kinds like this. Some possibilities that apply to particular kinds of natural kinds are that we ordinarily treat *M* and *N* as being of the same "substance" kind if we accept [*M* is made of the same stuff as *N*], and we ordinarily treat them as being of the same "species" kind if we accept [*M* can have fertile offspring with *N*] or [*M* and *N* can have fertile offspring with the same things]. But other possible tests are generally applicable to virtually all natural-kind categorization: [*M* obeys enough of the (basic, important) laws of *N*], or [*M* shares enough (basic, important) explanations (causes, effects) with *N*].<24> My psychological claim is that virtually no concepts (at least in minds free of special philosophical training) are *completely* free of such tests. This is because we should not and ordinarily do not establish *a priori* which of our concepts are natural-kind concepts and which are not.

Consider an apparently non-natural-kind concept such as [sofa]. Although the procedure would be hard to implement, we might investigate my claim by checking how people (without special philosophical training or other brain damage, and on reflection) would react if they came to believe, say, that the things we've been calling 'sofas'--and buying from furniture stores, sitting on in our living rooms, etc.--are not made in factories but instead are (extremely cooperative) living organisms. Would these creatures be sofas?<25> What if the same were true of things we've been calling 'toaster ovens'? What if we discovered that the alleged toaster-ovens are the *children* of the alleged sofas, and the former transform into the latter as caterpillars transform into butterflies? Are we prepared in principle to discover empirically that toaster ovens are (baby) sofas, even though they have virtually none of the alleged (stereotypical or classically "defining") features of sofas? If so--and I at least think so--it may be because we have (actual or idealized) natural-kind-*ish* influences even on categories that *in fact* we take not to be natural kinds.

The *Mars* and *Venus* cases give some reason to believe this claim even for [bachelor]. Like metalinguistic tests, in certain contexts, natural-kind tests for [bachelor] and [sofa] are

psychologically stronger than stereotype-based or allegedly classical tests. A systematic *pattern* emerges which threatens the claim that there are classical definitions with maximal psychological strength, and so which undermines (actual or idealized) maximal strength as a criterion for distinguishing nonholistic classical definitions from a much wider set of stereotypical commitments. Pending a better criterion, I will speak of allegedly classical tests as components of stereotypes and stereotypical tests. [Male] and [unmarried] are parts of some very strong but not maximally strong stereotype-based tests for [bachelor].<26>

So far we have considered contexts in which metalinguistic and natural-kind tests are stronger than stereotype tests. Are either metalinguistic or natural-kind tests maximally strong? There are cases in which stereotypes "rise again" as strong as or stronger than *both* metalinguistic and natural-kind tests. Sometimes we are *torn* between favoring the stereotype-based tests and deferring to experts or natural-kind considerations. Many people waver about whether whales are fish, given that they fit stereotypes of fish but are not of the same natural kind as most alleged fish, and so are not called 'fish' by experts. I've heard that peanuts turn out to be more like most beans than like most nuts, and I feel an impulse to admit that peanuts aren't nuts, immediately followed by the thought that I'll be *damned* if peanuts aren't nuts. Even with categories such as electron and gold, for which one is most strongly inclined to defer to experts and natural kinds, there are contexts in which entrenched stereotypes are too strong to squelch completely, and surface to control categorization. Maybe these cases can be idealized away as careless or uncircumspective; maybe not. But there are sufficiently abnormal contexts in which, even for categories in which one is most strongly inclined to defer, one favors one's lingering stereotypes *on reflection*. Consider a case due to Peter Unger; call it *Double Illusion*:

> [S]uppose that entities of the sort that human beings have considered to be cats . . . have been secreting a substance which has hidden from us . . . their 'true natures.' . . . Each alleged cat, we suppose, has just the relevant objective properties that we have been attributing to those entities we have taken to be *dogs*! . . . [Furthermore,] entities of the sort that humans have regarded as dogs have been secreting another substance . . . [and] beneath the appearances affected by the alleged canines are just the objective properties we have so long been attributing to those entities regarded by us as *cats*!<27>

For example, we have normally *thought* that the things we have normally called 'dogs' were barking and chasing cars, but they were *really*, beneath the appearances, purring and chasing mice. And the alleged cats, which have seemed to purr and chase mice, have really been barking and chasing cars. Now, which entities, on reflection, should we take to be cats (if any) and which should we take to be dogs (if any)? Unger reports that the typical considered response to this example is to say that the alleged cats are actually dogs, and the alleged dogs are cats. This seems to be a case, then, in which one's stereotype tests for cathood influence categorization more than one's [thing of the same natural kind as enough alleged-cats] or [thing called 'cat' by experts] tests. We keep the connection between [cat] and the stereotypical [purring mouse chaser] test, rather than keeping the connection between [cat] and the natural-kind [alleged-cat] or metalinguistic [called 'cat'] tests. We are not maximally resistant to rejecting even our strongest metalinguistic and natural-kind tests, and we are not maximally resistant to imagining in detail how they could go wrong.<28>

If neither stereotype tests, alleged classical definitions, metalinguistic tests, nor natural-kind tests are (actually or ideally) maximally strong, it is not promising for inferential role semantics to use maximal strength in order to draw a constitutive/collateral distinction and avoid meaning holism. But a very similar line of thought applies to comparative (rather than maximal) strength. The upshot of the discussion of maximal strength is not simply that different representations vary according to which tests are strongest, but more interestingly that different (imagined) contexts of use for a *single* representation vary according to which tests are strongest. In some contexts [cat]'s metalinguistic or natural-kind tests trump its stereotype (including classical) tests, but in contexts such as *Double Illusion* [cat]'s stereotype tests win the competition. In some contexts [bachelor]'s stereotype (including classical) tests trump its metalinguistic and stereotype tests, but in contexts such as *Wise One*, *Mars*, and *Venus* the reverse holds. Representations not only lack tests that are maximally strong, but also lack tests that are comparatively strong*est* in the face of arbitrary evidence. If so, no test for a representation is comparatively strongest independently of imagined context of use.

I think this presents a problem for anyone inclined to use comparative strength to draw a constitutive/collateral distinction. Psychosemantic meaning is supposed to help specify and explain generalizations about how representations (actually or ideally) are disposed to behave. To play this role, a representation's meaning should be something it "brings" to various circumstances of use, not something that varies according to circumstance. (We can imagine circumstances in which a representation completely and arbitrarily changes its behavior, and completely changes its meaning, but these are precisely the circumstances in which the meaning fails to explain the behavior.) Since comparative strength varies among contexts that do *not* involve arbitrary or complete changes in behavior, the constitutive/collateral distinction should not be drawn in terms of comparative strength, or else meaning would vary among contexts in which it should be stable.

In this section and the previous one I have argued against certain attempts, on behalf of inferential role semantics, to avoid meaning holism. I have given reasons for believing that our (actual or idealized) inferential dispositions with respect to our concepts do not generate a reasonable strength-based account of meaning-relevance. In particular, I think we may well have (i) a fundamental disposition inferentially to relate a concept to virtually every other concept (in *some* way--see note <12>), (ii) a fundamental disposition not to treat any of a concept's inferential connections as maximally strong, and (iii) a fundamental disposition not to treat any of a concept's inferential connections as comparatively strongest in all circumstances. Perhaps, nevertheless, there is some other psychologically nonarbitrary way for IRS to resist holism. But to put the point as cautiously as I can, if we want IRS, we should seek a way to defend one *in case* it leads to meaning holism, and *in case* no suitable restriction on meaning-relevant tests drops from the sky. That is what I tried to provide in section 2, by appeal to multiple meanings.

NOTES

<*> I would like to thank Ned Block, Robert Cummins, Michael Devitt, and Georges Rey for their reactions to these proposals and arguments in various disguises.

<1> A system is a group of representations poised to work together as (parts of) reasons, e.g., as (co)premises or (co)conclusions of inference. Natural languages and individual minds (at a time) may

serve as familiar examples, although minds perhaps divide into smaller systems whose representations are inferentially shielded from one another--see Jerry Fodor, *The Modularity of Mind* (Cambridge: MIT Press, 1983).

<2> Since I will often need to refer to specific (alleged) mental representations in giving examples, it will help to establish some descriptive conventions. I will refer to mental representations using formulae enclosed within square "mental quotation mark" braces, although I do not assume that mental representations must have all of the typical properties of formulae--internal syntactic structure, ability to be written and copied, etc. (Thus my terms ‘representations’ and ‘ideas’ are catch-alls for items in a language of thought, in a connectionist network, in a system of mental imagery, etc.) For convenience, I will normally use formulae of English, as in [snow is white]. When I use a linguistic representation in braces, I mean to specify a mental representation that distinctively "governs" at least some of one’s uses of the linguistic representation. In expressing a certain thought, which involves the representation [snow is white], one is likely to utter the words ‘snow is white.’ On occasion, I mean only that the mental representation is in some specified way similar to those which people may use to govern their linguistic representations.

<3> There are corresponding theoretical interests in the meaning of nonmental representations, especially "idiolect" meaning or the alleged meaning one’s words have "to one" independently of anyone else’s use or understanding of these words. But concern with mental representations is more likely to yield interesting generalizations, since the behavior of mental representations is plausibly more regular than that of linguistic representations. Perceptions and thoughts about redness are much more likely to bear lawful or other systematic relations to red things (and to other mental states) than utterances about redness are to bear such relations to red things (and to other utterances).

It is common to reserve the word ‘content’ for mental meaning, and the word ‘meaning’ for linguistic content, but I will speak unreservedly. Despite the focus on mental rather than linguistic meaning, psychosemantics does not necessarily "change the subject" of meaning or adopt a new sort of meaning, such as narrow content. It is a perfectly possible for psychosemantic interests to converge with other interests in meaning. However, I will neither assume nor deny this at the outset. I regret that space does not allow a proper treatment of the many options for extending the present discussion to public language.

<4> For defense of inferential role semantics, see especially Ned Block, "Advertisement for a Semantics for Psychology," in *Midwest Studies in Philosophy*, vol. X, 1986, pp. 615-78. IRS is also sometimes called ‘conceptual,’ ‘functional,’ ‘psychological,’ or ‘causal’ role semantics.

<5> Any inferential role theory of meaning needs a noncircular--and so, roughly, nonsemantic--account of *which* causal relations among representations count as "inferential" for purposes of the theory. Is the "associative" causal connection between [salt] and [pepper] inferential? How about the causal relation between representations early in the visual system and perceptual beliefs? I hope my discussion is abstract enough to be consistent with any natural answers to such questions.

<6> Fodor defends a referential psychosemantic theory, primarily to avoid holism, in *Psychosemantics* (Cambridge: MIT Press, 1987), and again in "Substitution Arguments and the Individuation of Beliefs," in *A Theory of Content* (Cambridge: MIT Press, 1990), pp. 161-76.

<7> See Fodor, *ibid*.

<8> I don't require this to be *causal* dependence, to keep compatibility with the many standard views according to which meanings don't have causal powers at all.

<9> The view that a token has more than one meaning seems to be an idea whose time has come. It has been independently suggested by Akeel Bilgrami, *Belief and Meaning: The Unity and Locality of Mental Content* (Cambridge: Blackwell, 1992), and Michael Devitt, "The Methodology of Naturalistic Semantics," *this JOURNAL*, volume XCI, number 10, October 1994, pp. 545-72. Devitt explicitly rejects holism, however, and Bilgrami seems to hold that the multiple meanings are not had all at once, but tend rather to shift with the interests of interpreters from one use of a token to another.

<10> In addition to the illustrations immediately following, I will describe several more such tests in sections 3 and 4. In "How to Be a Meaning Atomist" (under review) I try to characterize the right notion of a unit not in terms of tests but in terms of certain inferential dispositions to substitute representations for one another.

<11> Let me try to address briefly two concerns about how units can "yield" meanings. First, what happens when the tests "come apart"--that is, determine different properties or classes of things? In this case Larry's [bird] has multiple referents (or multiple conditions for reference) as well as multiple meanings. Second, how can the representations in the units for [bird] yield a meaning for [bird], if meaning holism is also true of them--and so ultimately their meaning depends on [bird] itself? In "How to Be a Meaning Atomist", *op. cit.*, I describe how MM reconciles meaning holism with "meaning atomism", the view that the meanings of all the representations in a system can be specified completely in terms of a relatively small set of semantically atomic (or simple, primitive, basic) representations in the system. A representation can have *a* primitive meaning--and so be a semantic atom--while also having *other* nonprimitive meanings--and so satisfy inferential role semantics and meaning holism. Units consisting entirely of atomic representations, then, can yield meanings which are independent of other representations in the system.

<12> A *complete* specification of the tests for a typical conceptual representation is likely to include virtually all of one's other representations. For starters, there are beliefs that yield very weak and normally "nondiagnostic" tests--[Mickey Mantle is allergic to birds], [birds are smaller than the Milky Way], etc. Also, for any potential belief *that p*, there are *some* contexts in which one may be disposed to use it to decide whether something is (say) a bird. Consider a context in which your only handy test for birdhood is my testimony, and your only handy test of my general reliability is whether or not you agree with my belief *that p*.

<13> MM also has this advantage over so-called "two-factor" versions of IRS (see, especially, Block, *op. cit.*). The basic idea of a two-factor theory is to identify the meaning of a mental representation *r* with a *pair* consisting of (i) *r*'s reference (i.e., *r*'s referent or reference condition), and (ii) all or part of *r*'s inferential role (i.e., the ways *r* is inferentially connected to other mental representations). Two-factor theories can provide for coarse-grained "synonymy" of the referential factor in meaning, but not for fine-grained stability, pending some nonholist breakthrough in the attempt to draw a constitutive/collateral distinction.

<14> See, e.g., Block, *op. cit.*, p. 629.

<15> Philosophers tend to use the word 'concept' *not* for mental representations themselves, but for their meanings. It can be easy for a philosopher to hear psychological claims about concepts, then, as claims directly about meaning. I will reserve the word for conceptual representations (meaning *bearers*) rather than risk speaking confusingly about the (philosophical) concepts expressed by (psychological) concepts.

<16> The most crucial early research is reviewed in Edward Smith and Douglas Medin, *Categories and Concepts* (Cambridge: Harvard University Press, 1981).

<17> In these examples I don't mean that [similar to] and [enough of] govern uses of 'similar to' and 'enough of.' Rather I mean [similar to] simply as a placeholder for whatever relation is defined by the correct psychological theory of how the thinker computes similarity to exemplars. Similarly, [enough of] might simply stand for whatever the right combinations of [F1,...,F*n*] are for the person to apply the representation in question.

<18> Although in this paper I will focus on strength as an initially attractive means for drawing a constitutive/collateral distinction, it is not the only available means. Another family of options involve the "depth" of a test, for example *primitive depth*, or degree of independence of the test from other tests, and *explanatory depth*, or degree of dependence of other tests on the test. The size test for [superlune] is shallow in that its existence depends on the existence of the mass test (and not vice versa). To assess this properly we would need to examine several variants. Must a test be maximally deep to be meaning-relevant, or is it enough to be comparatively deepest? Is the relevant kind of dependence generative (a matter of one test actually having been caused by another) or sustaining (a matter of one test's ceasing to exist were another test to cease to exist)? We will have plenty to do attending only to strength. I think many of my arguments against strength are relevant to depth, but there is no space to pursue this here. Roughly, I think that depth does a little better than strength in avoiding *extreme* meaning holism, but still leaves enough holism to threaten stability and to call for a move to multiple meanings (as in section 2).

<19> Peter Unger engages in some preliminary bridge-building in "The Causal Theory of Reference," *Philosophical Studies*, vol. 43, 1983, pp. 1-45.

<20> Georges Rey suggests this strategy in "Idealized Conceptual Roles," *Philosophy and Phenomenological Research*, vol. 53, 1993, pp. 647-52.

<21> *The Nature of Selection* (Cambridge: MIT Press, 1984), p. 66. Sober credits the case to Philip Kitcher.

<22> See "The Meaning of 'Meaning,'" in *Mind, Language, and Reality* (Cambridge: Cambridge University Press, 1975), pp. 215-272. Tyler Burge further illustrates the role of linguistic communities in "Individualism and the Mental," in *Midwest Studies in Philosophy*, vol. 4, 1979, pp. 73-121.

<23> Putnam discusses robot cats in several early papers, and *op. cit.*, pp. 243-4.

<24> Although concept psychologists are beginning to investigate these possibilities, very little work has been done relative to the amount of work on stereotypical representations. Perhaps this is because, unlike stereotypical or metalinguistic tests, natural-kind tests are very infrequently used in normal categorization. While we often key on *particular* behaviors or composition in categorization, the natural-kind tests in question are more *abstract*: we think cats behave or are made in the same way as enough samples, *whatever that way is*.

<25> The case thus far resembles one credited to Rogers Albritton by Putnam, *op. cit*, pp. 242-3.

<26> Once grasped, the pattern can be extended with only moderate difficulty to other terms paradigmatically "definable" in a nonholist manner. We normally say that *A* is a grandfather of *B* iff *A* is a father of a parent of *B*. End of story? A bizarre kind of counterexample may be sobering. Thanks to science fiction, it is (rightly or wrongly) part of the public imagination that a man might travel backwards in time and (help to?) father himself. What is interesting is that we do *not* think that if a man *were* somehow his own father, he would thereby be his own *grand*father (and greatgrandfather, etc.). But if 'grandfather' is defined as 'father of a parent,' this should *follow*--if *B* is a father of *B*, then *B* is a father of a parent (namely, *B*) of *B*. But I think we are inclined to block this inference, even on reflection. Of course, we can set off on the usual philosophical wild-goose chase, attempting to modify the definition in the face of similar examples. (Maybe the parent of *B* and the father of the parent of *B* must be nonidentical. But what if they are identical via reincarnation ...?) As the definitions and counterexamples get more and more complicated, I think what would (rightly) be governing our reactions to each case is the natural-kind-ish and metalinguistic-ish pull to find something (whatever it is, perhaps hidden) that unifies enough of the familiar things we've been calling 'grandfathers'--mine, yours, etc. The exemplars (and their holistic stereotypical and hidden properties) are what we have left when we lose our overconfidence in our nonholist definitions.

<27> Unger, *op. cit*., p. 9.

<28> Should reactions to the *Double Illusion* case be idealized away? Not obviously. They cohere well with explanatorily promising idealizations. People categorize by finding some sort of "best fits" that *potentially* take into account *any* of their beliefs about a category, so that under idealization people have a rational competence for categorization *all things considered seriously*. Why ever *should* we do otherwise, idealizing away from deficiencies in memory, time, and other limited resources? Of course, I grant that *some* of our dispositions are the results of slips and other interfering "performance" factors. But I expect that even if we idealize these away as irrelevant to psychosemantic meaning, for each category psychologists will find "messy" underlying dispositions of the sort I have illustrated.