Theoretical Terms without Analytic Truths

Michael Strevens

Philosophical Studies, 160:167–190, 2012

Abstract

When new theoretical terms are introduced into scientific discourse, prevailing accounts imply, analytic or semantic truths come along with them, by way of either definitions or reference-fixing descriptions. But there appear to be few or no analytic truths in scientific theory, which suggests that the prevailing accounts are mistaken. This paper looks to research on the psychology of natural kind concepts to suggest a new account of the introduction of theoretical terms that avoids both definition and reference-fixing description. At the core of the account is a novel psychological process that I call *introjection*.

Where do new theoretical terms come from? Either singly or as members of a terminological family, they are introduced into science by neologistic acts that bestow upon them whatever kind of semantic significance is necessary to do scientific work. This paper is about the nature of those neologistic acts, about the sort of semantic significance with which new terms are imbued as a consequence of the acts, and therefore about the sort of semantic significance that a term must possess in order to function fruitfully in scientific inquiry.

I will focus on one question in particular: must a neologistic act—at least, one that coins a new theoretical term—create analytic truths involving the predicate or name in question? (By an analytic truth, I mean a proposition that is both true and known to be so a priori in virtue of the semantic properties of its constituents, that is, give or take some connotations, what is often called

a semantic truth, a conceptual truth, or a stipulative truth. Perhaps not all semantic truths are a priori; it is the a priori ones, however, that concern me here, and so I build aprioricity explicitly into analyticity.)

Past and present theories of theoretical terms have tended to answer in the affirmative: a new term always makes its scientific debut with an entourage of analytic truths. But there are reasons to think that science as we practice it contains few or no analytic truths about theoretical entities. Perhaps, then, we need a new account of theoretical terms, and most of all, a new account of the acts by which these terms make their way into scientific discourse. I will propose such an account.

1. Origins and Analyticity

Let me begin by surveying some recent ideas on the semantics of theoretical terms, in order to show you how a wide array of theories end up committed to the existence of analytic truths created by the neologistic acts that introduce new terms. One caveat: ignoring Quine's (1951) famous argument, I assume that a principled distinction can be drawn between analytic and empirical truths.

If there is a standard account of theoretical terms, it is this: a new theoretical term is introduced into scientific discourse by way of an explicit definition, expressed using vocabulary that is either primitive or that has previously been well defined. On the standard view, the term 'pressure' might have been introduced by defining pressure as force exerted per unit area, while 'fitness' might have been introduced by defining an organism's fitness as its expected number of offspring.

The definitions themselves are stipulations, and so create analytic truths, most notably the elements of the definition itself: it becomes an analytic truth, for example, that pressure is a kind of force or that two people with equal expected numbers of children are equally fit, even if they have different expected numbers of grandchildren. Such truths cannot reasonably be

disputed: if you deny that pressure is a force, either you do not understand the meaning of the term or you are using a different term that is orthographically indistinguishable from 'pressure'—different because it has a different definition, a definition that does not entail pressure's forcehood. You might, of course, suggest a revision to the definition on the grounds of conceptual economy or simplicity; you are not then disagreeing with the original definer about pressure, however, but rather on whether the suite of concepts that best organizes the relevant body of scientific knowledge contains a term meaning "force per area unit". In a sense, you are arguing about the clerical, not the empirical, virtues of the stipulated propositions.

Several notable amendments have been made to the standard view. First, it has been allowed that terms may be introduced by implicit as well as explicit definition. An implicit definition is a system of sentences connecting new and older terms that do not use the language of definition explicitly but that are stipulated to be true, and jointly sufficient to give the terms the entirety of their meaning, nevertheless.

Second, Carnap introduced the notion of a partially interpreted term, a term that is given a meaning by what is in effect a definition part—a necessary but insufficient criterion for falling under the term, such as "Bachelors are males" or "Squares are four-sided polygons".[1]

Third, Lewis (1970, 446) suggested that a definition might have the structure of what psychologists call a prototype, as opposed to a list of necessary and sufficient conditions: rather than the definition of T specifying that Ts must satisfy every one of a list of requirements, it might specify that Ts must satisfy sufficiently many of the requirements on a list.[2]

Fourth and finally, post-empiricists committed to the theory-ladenness of observation have suggested that even what appear to be primitives may

1. On the history, interpretation, and criticism of the notion of partial interpretation, see Suppe (1977, §4C). Suppe also provides a useful discussion of the standard view.

2. Lewis is more concerned with reference than definition, but he is clear that his reference-determining descriptions constitute definitions of the terms in question.

be definitionally entangled with theoretical terms. As a consequence, the meaning of the primitives may change as new terms are defined. This is one way, at least, to understand Feyerabend (1962).

These amendments all share with the standard view the notion that theoretical terms are introduced by making stipulations, and that their introduction is therefore accompanied by the creation of stipulative truths, hence analytic truths. Perhaps this is least clear with Lewis's "prototype" view; however, it is not hard to see. If T is defined so that x is a T just in case x satisfies any two out of the three conditions P, Q, and R, then it is an analytic truth that all Ts satisfy at least two of P, Q, and R. To maintain that some Ts fail to satisfy both P and Q, for example, would be either to misunderstand the meaning of 'T' or to make a logical error.

The development of the causal-historical theory of reference (Kripke 1980) inspired a significant departure from the standard theory. The reference of a natural kind term such as 'gold' could not, Kripke argued, be determined by the kinds of descriptions we associate with gold, because any or all of these descriptions could turn out to be false of the extension of 'gold'. It follows, among other things, that none of the descriptions we associate with gold can be true of it by stipulation. Kripke went on to suggest that the reference of natural kind terms is fixed by an act of baptism, or something with much the same semantic consequences, and a chain of causal links connecting later uses of the term to that baptism. The baptism includes a stipulation of a sort, note—"I hereby introduce the term 'gold' to refer to that kind of stuff" (pointing to some stuff)—but not one that makes a stipulative truth out of any of our practical or theoretical knowledge about the substance in question.

Kripkean semantics has been applied to theoretical terms by philosophers in defense of scientific realism (Hardin and Rosenberg 1982). The realists propose, first, that the only semantics a theoretical term requires is an extension; thus, there is no need of a definition to supply a term with a Fregean sense or

any such thing.[3] They then adopt a causal-historical account of the reference of theoretical terms. Kripkean reference-fixing serves the realist cause because it allows for considerable semantic continuity between old scientific theories and the new theories that replace them even when the theories have rather different central tenets: the 'gravity' of Newton and the 'gravity' of Einstein, for example, can be understood to be one and the same term, on the grounds that they have the same reference, despite the fact that any likely Newtonian definition of gravity would fail to pick out Einsteinian gravity.[4]

The recent consensus among realists and other writers is, however, that the causal-historical theory as stated is inadequate as an account of the reference of theoretical terms, because the initial baptism required by the causal-historical account cannot succeed in picking out unobservable entities without some sort of descriptive supplement. It is no use to point to a proton saying "By 'proton' I mean one of those"; if you are pointing to a proton you are also pointing to a quark, and likely pointing to electrons, neutrons, atoms, and various pieces of laboratory equipment as well. Something must disambiguate your act of ostension.[5]

In many and perhaps all cases the disambiguating factor will, it seems, be a specification of a theoretical property of the entity in question, such as its playing a distinctive causal or explanatory role in the theory. Psillos (1999), for example, suggests that each theoretical term is associated with what he calls a "core causal description", specifying the causal properties in virtue of which the putative entity explains what it does by the lights of the theory. The core description does most of the work in picking out a theoretical term's

3. I will not distinguish in this paper between a term's extension in the sense of the set of actual objects falling under the term, on the one hand, and in the sense of the property picked out by the term, on the other. An extension or reference in the latter sense determines what falls under the term in any possible world, hence the term's intension.

4. Psillos (1999) provides a useful overview of the role of Kripkean semantics in the realism debate.

5. The same objection has been leveled against Kripkean reference-fixing even for natural kind terms, but that will not concern me here.

referent; the causal-historical chain's role is subsidiary.

There are many variants on what is now usually called "causal descriptivism" (Enç 1976; Sterelny 1983; Kroon 1987). Some are purely descriptivist, causal only in the sense that the reference-fixing descriptions in question specify a causal role. Some, like Psillos's, use descriptions to help fix reference but also rely on causal connections. Either way, the use of descriptions to fix reference will create analytic truths. If I point to an electrical current and say "By 'electron' I mean the things whose flow through wires in experimental setups like these creates such and such an effect", then it becomes an analytic truth that electrons, if there are such things, have such and such an effect in such and such circumstances, because it is a logical consequence of the nature of reference (according to causal descriptivism) along with the details of the reference-determining act that 'electron' can have an extension only if electrons flowing through a wire in the circumstances in question have the effect in question.[6] Thus the semantics of 'electron' entails either that there are no electrons—if the reference-fixing act fails—or that electrons have the properties specified by the reference-fixing description.

If it is allowed that claims about electrons presuppose their existence, then, causal descriptivism entails the existence of analytic truths of the form "Electrons are negatively charged". Without the presupposition of existence, the analytic truth will have the form "Electrons, if they exist, are negatively charged". Either variant will create the same difficulties. (A more liberal version of causal descriptivism, which allows descriptions to pick out entities that they to some extent mischaracterize, will be considered in the next section.)

What I have considered so far are only a few rather simple accounts of the introduction of theoretical terms (and there are some sophistications to

6. The "circumstances in question" might be determined by an act of ostension—pointing to a particular experimental setup—along with some understood standard for determining what counts as other setups "like this". A further description seems to be required to decide the standard in question; it is this sort of consideration that has pushed incipient Kripkeans back toward a descriptivist account of theoretical terms.

come in the next section), but I conjecture that the lessons learned generalize: any introduction of a theoretical term that involves an intention on the part of the introducer to give the term a semantics, such as "By this term I mean ____" or "This term refers to ____", will create corresponding analytic truths. Which should raise the question: how might a term acquire a semantics, if not by way of such an intention? The ultimate aim of this paper is to give that question an answer.

2. Analytic Truths: The Good and the Bad

There is much to be said for analytic truths, most of all when they come in the form of complete definitions. Definitions are a wonderful way to organize and regiment scientific language. As many writers have noted, a definition provides a term with both sense and reference. The sense of a term is constituted by the propositions that make up the definition itself. A scientist understands the term just in case they grasp these propositions, and grasp the fact that they constitute the definition. Two scientists use a term in the same way if they associate it with the same definition. The extension of a term is whatever in the world satisfies the definition. If nothing fits the definition, then the term fails to refer. Finally, definitions enable a straightforward story about the introduction of new theoretical terms.

Were science to be organized or reorganized so as to maximize its congeniality toward the orderly, philosophical mind, theoretical terms would surely get their semantic significance by way of definitions (Carnap 1937). But there is good reason to think that, whatever the merits of definition-driven semantics, it plays at most a minor role in science as we know it.

The good reason is supplied by an observation familiar from Quine (1951, 1963): no, or almost no, truth about the entities picked out by our science's theoretical terms seems to be immune to empirical disconfirmation or refutation. (Quine concluded that all scientific propositions were "gray", suspended somewhere between the analytic and the empirical, but his evidence supports

equally well the conclusion that they are "black", or purely empirical—which is what I suppose here.)

I should allow for some exceptions to the rule. The proposition *The term 'phlogiston', if it refers to anything, refers to phlogiston* is presumably empirically untouchable. It is not clear that a proposition such as *Phlogiston, if it exists, is a substance* could be disconfirmed (though nor is it clearly immune to disconfirmation). To take a different kind of example, the analyticity of Chalmers's (2006) "primary intensions", which determine mappings from possible worlds to extensions for theoretical and other terms, is in principle compatible with the confirmation or disconfirmation of any normal scientific hypothesis; the primary intensions' analytic status seems, then, to be consistent with the observed nature of science.[7]

What do appear clearly to be subject to disconfirmation are specific statements about the causal powers of a property or substance. Both the standard and the descriptivist-Kripkean accounts of theoretical terms imply that some such "hypothesis-like" statements are analytic. It is this implication that conflicts directly with scientific practice.

Let me consider several ways to defend an analytic truth–implying view of theoretical-term introduction. First, could what appear to be revisions of a hypothesis in response to unfavorable evidence in fact be revisions of a definition to enable a reorganization of the other, empirical parts of a theory in response to the evidence? Clearly, definitions may be altered in this way in principle. What will distinguish the two operations is the kind of argument advanced for the revision. Suppose that an evolutionary biologist proposes a change in some proposition concerning fitness. If the proposition is empirical, the argument will have the following form: previously, we thought such and such about fitness, but there are reasons to think that this is wrong; here is the

7. Of course, a certain kind of primary intension *could* entail what appears to be a substantive, empirical claim about the world; indeed, Chalmers's argument for dualism turns on just such an implication.

correct story. If the proposition is an analytic truth, talk of right and wrong is inapt; instead, the argument will concern the structural benefits of a new definition, considered as a reorganization rather than a revision of a theory's empirical content.

In fact, talk of structural or organizational benefits is as rare in science as it is common in mathematics. Scientists, when they propose changes to a theory—for example, arguing whether fitness should be understood to incorporate variance in, as well as expected number of, offspring (Gillespie 1977)—almost always talk in terms of right versus wrong, not economical versus prolix or convenient versus awkward.

A second defense: might scientists be changing meanings without understanding the true nature of their revisions? Perhaps, if reference-fixing is a complex matter, the analytic truths introduced alongside a new theoretical term will be baroque or obscure. Researchers in the thick of inquiry might lose track of what holds in virtue of semantic facts alone and mistakenly treat such truths as empirical.

This is, of course, a paradigm of ad hockery: explain away the absence of the phenomenon predicted by your hypothesis with another hypothesis tailored to make that phenomenon invisible. The suggestion is not entirely unreasonable, but it surely shows that it may be worth considering alternatives to accounts of theoretical neologism that generate analyticities.

What if (the third defense) the analytic truths created by constructing new theoretical terms are not large, sweeping theoretical propositions but more local and particular states of affairs? Suppose, for example (and unhistorically), that the term 'electron' is introduced by a description of the form "Let 'electron' pick out the particles, an excess of which is responsible for the negative charge of these particular oil drops".[8] It becomes an analytic truth that the long-

8. As written, the description seems to presuppose a certain explanation for the negative charge of the drops. It would be more in line with the descriptivist approach to eliminate all presupposition using appropriate quantification, as in Russell's theory of descriptions or Ramsey's eponymous sentences.

vanished oil drops owed their negative charge to a surfeit of electrons, but since this is not itself a matter of any great scientific importance, its immunity to empirical refutation will have no impact on the epistemic conduct of science in later ages.

It is not true, however, that the analyticity, local and historical though it may be, has no implications for later theorists. There are some aspects of physical theory that, were they amended or abandoned, would call into question whether the negative charge on the baptismal oil drops was in fact due to an electron overload (as if, for example, it later turned out that some other kind of negatively charged particle, such as the muon, was generated in abundance by the oil-drop apparatus).[9] If it is an analytic truth that the reference-fixing description of 'electron' is true of the electrons, then these theoretical tenets must themselves be immune to empirical disconfirmation. I have no proof that a description introducing a theoretical term must always constrain theory in this way—that it must always incorporate some inferentially important theoretical content—but the trend to causal descriptivism is due in large part to the absence of philosophically plausible alternatives.

A fourth defense of causal descriptivism liberalizes its apparatus to a certain extent. So far, I have assumed that causal-descriptivist reference-fixing can succeed only if the reference-fixing description is satisfied. Some well-known work in the philosophy of language suggests, however, that in certain circumstances, a description can succeed in picking out a referent without being true of that referent (Donnellan 1966; Kripke 1977). To use Donnellan's example: "the man drinking the martini" might succeed in picking out a certain salient individual holding a martini glass even if the glass contains only water.

Suppose that a physicist points to an oil drop suspended in an electrical

9. It would also have to turn out that such muons are unusually long-lived, and so on. These discoveries, though conceivable, are certainly not, given what we already know, particularly likely.

field and says "By 'electron' I mean the kind of particle, a surfeit of which is responsible for the negative charge in virtue of which that oil drop is suspended in the field". Suppose also that the oil drop, though negatively charged with a surfeit of electrons, is suspended not because of the charge—which is too weak for the job—but by a freak laboratory zephyr. Would not another physicist, conversant with all the facts, consider their colleague to have successfully picked out the electrons on the drop? Might this not mean that the extension of 'electron' is successfully fixed even in the absence of so charitable and well-informed a witness?[10]

Affirmative answers to these questions imply the following variant on causal descriptivism: reference-fixing by description attaches a term to whatever a charitable, well-informed observer would take to be the intended reference of the description. Reference-fixing fails, then, only if the putative observer would come up empty-handed in their best attempt to make sense of the description.

Such failure is perhaps rare when the describer directly perceives the intended referent, but less so when their access to the referent goes by way of a chain of beliefs that might be deeply mistaken. In the case of the oil drop, for example, even a charitable observer could not say that the description succeeded in picking out electrons as opposed to, say, muons if there was no surfeit of electrons and so no negative charge on the oil drop (remembering that the description is deployed before the concept of an electron, let alone that of a muon, has entered physics).

Now observe that even on the charitable theory, reference-fixing creates an analytic truth: electrons, if they exist at all, must satisfy the reference-fixing description well enough that a charitable, well-informed observer would consider them to be the determinate target of the description. Such truths, though weak, put strong constraints on theoretical change. As already remarked, it

10. For related reasons, Kitcher (1978) argues that Priestley's term 'dephlogisticated air' in some circumstances referred to oxygen.

seems perfectly conceivable that later discoveries in physics would show that electrons, though responsible for most of the phenomena subsequently attributed to them, did not create a negative charge on the oil drop in question. In such a case, even the most sympathetic observer would judge the act of reference-fixing to fail; thus, it is possible that the supposed analytic truth turns out to be false. The liberal form of causal descriptivism that implies the irrefutability of the "truth" is thereby itself proved false.

My tentative conclusion is that most or all of our science's theoretical terms have been introduced without either reference-fixing descriptions or definitions.

What are the alternatives? There are a number of possibilities. You might, like Quine, deny the coherence of a distinction between analytic and empirical (or other synthetic) truths, giving an account of evidence-driven theoretical revision that steers between those poles. You might opt for a notion of "graded analyticity", on which a truth can have a definition-like aspect while not removing itself entirely from the empirical firing line. You might try to develop a Kripkean account of reference on which the baptismal intentions that help to fix reference somehow avoid apotheosis as analytic truths. You might argue that theoretical terms are introduced by specifying "primary intensions" in Chalmers's (2006) sense. (As noted above, such a specification creates an analytic truth, but one potentially so flexible that it is compatible with the empirical refutation of any proposition having the character of a normal scientific hypothesis.) All of these suggestions are worthy of further investigation; each brings its own formidable problems. Rather than pausing to assess their prospects, however, let me make my own proposal, based on an account of concept formation provoked by recent work on the psychology of concepts of natural kinds.

3. The Standard Theory of Concept Acquisition

Following the psychological literature, I will work with a narrow notion of natural kind, focusing on chemical substances and biological species (or more exactly, biological categories at the "folk genus" level), such as gold, water, swan, and tiger. I will not address the question whether the theories of concepts described below apply to the broader range of categories that philosophers have called natural kinds. In what follows, then, a natural kind concept is simply a concept picking out either a chemical substance or a folk genus or similar biological category.

There is, in psychology, a default view about the acquisition of concepts of any sort: new concepts are acquired by acquiring definitions. Such definitions are either learned or built from scratch, depending on whether the concept itself is learned or built from scratch. When you are learning your first language or taking your first class in quantum mechanics, you are learning pre-existing concepts—the concepts of a linguistic or scientific community—by learning other people's definitions. In other circumstances, as for example when on a solitary mission of exploration you see some new kind of animal, you build your own concept by building your own definition.[11]

Call this the *standard view* of concept acquisition, since it is held by the majority of cognitive psychologists who have any definite view of concept acquisition at all. (Many do not.) The parallel with the standard view of theoretical terms is obvious.

The standard view is compatible with almost any theory of concepts. Let me give you three examples.

On the classical theory of concepts, which is attributed to Locke and earlier

11. Fodor (2008) claims that psychology's default theory of acquisition is one of hypothesis formation and testing. For supervised learning—learning another person's concept—he and I are both right: on the default view, other people's definitions are learned by forming and testing hypotheses about those definitions' contents. For unsupervised learning, we cannot both be right. I do not understand how hypothesis formation and testing even *could* be a theory of unsupervised concept acquisition, but perhaps that is Fodor's point.

writers still, a concept is a definition. In particular, a concept of a natural kind is a definition cast in terms of what might be called the characteristic properties of that kind. Locke himself suggested that the swan concept might be nothing over and above the concept of a white, red-beaked, trumpeting thing with the power of swimming, or in other words, that swanhood is simply defined as the property of being a white, red-beaked, trumpeting thing with the power of swimming. On this theory, the standard view of concept acquisition is mandatory: to acquire a concept is to acquire a definition.

On the prototype theory of concepts, a concept might be thought of as a description much like the category-defining description found in the classical view. The swan concept, for example, might be a list of properties no different from Locke's: white, red-beaked, trumpeting, swimming. The way in which the description—the "swan prototype"—is used to determine category membership is, however, non-classical. Whereas on the classical account, conforming to the description is necessary and sufficient for category membership, on the prototype account, an organism is considered a member of the category if it fits the description well enough, and better than it fits the description corresponding to any "rival" category. Something may be counted as a swan, then, because it has *most* of the properties mentioned in the description, or even if it has *some* of the properties on the list if those properties are considered important enough. As this example implies, different aspects of the description may be weighted differently for the purposes of determining goodness of fit, and many further refinements and variations are possible (Murphy 2002).

How are prototype concepts acquired? On the standard view of acquisition, by formulating a mental definition: *I hereby define a "swan" to be any organism that well enough fits the swan prototype* (where as remarked above, there may be a sophisticated algorithm for determining goodness of fit). Let me break down this acquisition process into several steps, focusing on the case in which the swan concept is not learned from others but introduced in response to

first-hand experience of a new kind.

Suppose, then, that the learner comes across a group of birds that fit no known prototype, birds that are white, red-beaked, habitual trumpeters, and so on. Here is what happens next:

1. The relevant properties of the novel birds are put together to form a new prototype.

2. A new, blank mental predicate 'swan' is minted.

3. A mental definition is made to link predicate and prototype: let 'swan' refer to just those things that well enough fit the prototype (and fit rival prototypes less well).

The prototype theory determines what kind of thing is linked to the new predicate, then, while the standard theory of acquisition determines the nature of the link—definitional.

On the psychological essentialist view, a natural kind concept is a theory, part metaphysical and part causal, of the kind in question. The metaphysical part posits the existence of an essence, an unobservable property that is necessary and sufficient for kind membership. The causal part asserts that the essence has a tendency to cause the kind's characteristic properties. The concept of a swan, for example, would be a theory with something like the following tenets: (a) there is a certain property *S* that all and only swans possess, and (b) *S* causes whiteness of the feathers, redness of the beak, habitual trumpeting, various features that enable swimming, and so on (Gelman 2003). A great virtue of the essentialist theory is its ability to explain various patterns of projection and categorization in normal humans' thinking about natural kinds, and in particular to account for aspects of these inferences that are characteristic of causal reasoning (Strevens 2000).

How is a psychological essentialist concept acquired? On the standard theory of acquisition, when an individual encounters swans for the first time, concept acquisition will proceed as follows:

1. The relevant properties of the novel birds are put together to form a new theory with the following tenets: (a) there exists an unobservable property *S* shared by the specimens whose observation prompts the creation of the new concept, and (b) *S* causes the birds' shared properties, namely, whiteness of the feathers, redness of the beak, and so on.

2. A new, blank mental predicate 'swan' is minted,

3. A mental definition is made to link predicate and theory: let 'swan' refer to just those things that possess the property *S*.

Note that the causal connections hypothesized in the first step, such as the proposition that the swan essence *S* causes whiteness, are not enshrined in the definition of 'swan'. The concept's possessor can therefore change their mind about swans' characteristic properties—they can change their minds about which properties are caused by the essence *S*—without revising the definition.

The standard view of concept acquisition, like the standard view of theoretical term introduction in science, posits a definition and so implies the existence of analytic truths. The nature of these truths varies with the theory of concepts:

1. On the classical view, it is an analytic truth that swans have the properties enumerated in the defining description—that swans are white, red-beaked, compulsive trumpeters, and so on. (Of course, this is true on the classical view however concepts are acquired.)

2. On the prototype view, it is not an analytic truth that swans are white, but it is an analytic truth that swans fit the swan prototype to some minimal degree.

3. On the essentialist view, the analyticity is subtler, yet still rather strong: swans must share a property causally responsible for their distinctive

observable characteristics (though as noted above, the identity of these characteristics is not specified in advance by the definition).

Here lies a problem that should have a familiar ring: our actual reasoning about swans seems not to respect the inviolability of a single one of these alleged analytic truths:

1. Upon coming across *Cygnus atratus* for the first time in Australia, Europeans concluded that some swans are black, as they could not if whiteness were built into their mental definition of 'swan'. They treated the whiteness of swans, in other words, as an empirical posit, subject to disconfirmation by the evidence, rather than as an analytic truth.

2. Just as it is possible for swans to lack one property thought to be characteristic of the species, they may lack many. Keil (1989) told children and adults a story in which a raccoon is, by various cosmetic and other treatments, made to have all the observable properties and behaviors of a skunk. Children and adults maintain that the animal remains a raccoon, as they could not if resemblance to the raccoon prototype were by definition necessary for membership of that species.

3. Essentialism, biologists tell us and proponents of psychological essentialism are quick to concede, is in fact false: there is no single property that is shared by all and only swans and that causes swans' characteristic observable properties. If the term 'swan' is defined in terms of an essence, then by definition a thing can be a swan only if it possesses such an essential property—the sort of property that no living being possesses. Thus there are no swans; the organisms to which we have been applying 'swan' time out of mind are in fact some other kind of thing. Of course we do not react to the refutation of essentialism in this way; we rather treat it as an empirical discovery about swans, just as early colonizers of Australia treated it as a discovery that some swans are black. Thus, we do not treat swan's possession of a characteristic

property–causing essence as an analytic truth, but rather as an empirical hypothesis.

On this showing, you might think that no hypothesis-like facts about swans whatsoever are treated by ordinary thinkers as analytic. That would be rash: surely there may be analytic truths about swans that are deeper, subtler, weaker than those whose existence is falsely implied by the standard account of concept acquisition conjoined with the theories of concepts surveyed above. Perhaps, then, we need to find a new theory of concepts, or to revise one of the known theories, in order to save the standard account of acquisition.

Then again, it is surely worth asking whether the standard account itself is mistaken. That will be my approach.

Begin at the end: what should a natural kind concept look like once acquisition is complete? The answer, I suggest, is that it should consist entirely of empirical hypotheses concerning the kind, and thus should neither contain nor imply analytic truths.[12]

One theory of natural kind concepts that satisfies this requirement is a view of my own called *causal minimalism* (Strevens 2000). You might think of causal minimalism as essentialism without the essences. Whereas on the essentialist view, a swan concept involves something like the following beliefs:

> There is a property *S* that is constitutive of swanhood: to be a swan is to have *S*
>
> *S* causes whiteness;
>
> *S* causes red-beakedness;

12. When I say that a concept "consists of" empirical hypotheses I mean that, whenever it is instantiated, it is realized by a set of empirical hypotheses (typically a different set at different times and in different minds). Thus, I do not mean to imply that a concept is *individuated by* the hypotheses of which it consists, or else any change in hypothesis would entail a change in the concept itself, rendering each of the hypotheses an analytic truth.

and so on, for all the other characteristic properties of swans, on the minimalist view the corresponding swan concept by contrast involves only the following beliefs:

> Something about swans causes whiteness;
>
> Something about swans causes red-beakedness;

and so on. Each of the causal hypotheses is, according to minimalism, a purely empirical hypothesis, subject to disconfirmation or refutation by new evidence. Equally, new hypotheses may be added to the list as new discoveries about swans are made.

A full presentation of causal minimalism would explain its empirical advantages—showing that it better explains the known facts about our reasoning about natural kinds than any other theory of natural kind concepts—and would also, as a matter of course, say more about the truth conditions for the causal hypotheses to which minimalism appeals, specifying what the locution "something about swans" amounts to (must it be something that all swans possess?) and spelling out the content of the hypotheses' implicit ceteris paribus hedges. You can find the empirical defense in Strevens (2000). An interpretation of "something about swans" is given in Strevens (2008a, §7.3) and Strevens (2008b); very roughly, what is implied is a counterfactually robust but not necessarily exceptionless relation between swanhood and the causal property in question. Finally, an interpretation of the ceteris paribus hedges is given in Strevens (forthcoming); the hedge restricts the scope of the claim to consequences of the operation of what you might colloquially refer to as "natural mechanisms", so that, for example, *Something about swans causes whiteness* makes a claim only about the end product of the "natural" swan coloration mechanism.

But none of this matters much for the purposes of the present paper; what is important here is simply that there is nothing to a minimalist natural kind concept above and beyond a set of empirical hypotheses that function much

like (in fact, exactly like) hypotheses you might find in a scientific theory.

Suppose, then, that there are independent reasons to accept the causal minimalist account of natural kind concepts, if only because it raises a highly relevant and interesting question: how do minimalist concepts get into the head? Since such concepts contain and imply no analytic truths about the kind in question, the acquisition process cannot involve a definition. The standard account cannot, then, explain how a minimalist concept is acquired. What is needed is a new theory of concept acquisition. To develop such a theory is my next goal.

4. The Introjective Theory of Concept Acquisition

While out for a walk one day, you come across a group of birds unlike any other birds you know of: they have blue feathers, purple beaks, a trombone-like call, and have no interest in or facility for swimming whatsoever. In the circumstances, it would be entirely reasonable to suppose that they are members of some previously unknown kind; call them *schwanns*. To think such thoughts, you would like to have a concept of this new kind, a schwann concept. Assuming that the causal minimalist theory of natural kind concepts is correct, how would such a concept be acquired?

A minimalist schwann concept is realized by a set of causal hypotheses—genuine empirical hypotheses, not analytic truths—of the following form: something about schwanns causes them to grow blue feathers; something about schwanns causes them to issue trombone-like calls; and so on. Any process that gets just these hypotheses and no others into the head will suffice for concept acquisition.

How easy is that? It is logically impossible, there is very good reason to suppose. For consider: to get such a hypothesis into the head is to take some kind of propositional attitude towards it, such as belief or a high level of credence (or, for that matter, a low level of credence). To take such an attitude, you must mentally represent the proposition in question, that is, you must

represent a proposition such as *Something about schwanns causes them to grow blue feathers*. To represent such a proposition, you must have a schwann concept. Thus, to acquire the concept you must already have the concept. Impossible.

Clearly, it is not anything peculiar to causal minimalism that creates this problem. Any theory according to which natural kind concepts are realized by empirical beliefs about the kind in question will face the same difficulty.

Let me generalize further. Suppose that the acquisition of any new concept, like the introduction of any new theoretical term, has two steps: first, a new blank, meaningless predicate is manufactured, and second, the predicate is connected with existing, meaningful terms. There seem to be two ways to make the connection: "use" and "mention".

The way of use: embed the blank predicate in some sentence otherwise constructed from pre-existing terms. Take some epistemic attitude toward the sentence, presumably belief or partial belief, or in other words, use the new concept in one or more new hypotheses. Problem: you have not yet acquired the concept, so you cannot use it. To put it another way, you cannot rationally take an epistemic attitude to the new sentence, because it contains a meaningless term, and so fails to express any determinate proposition. Or in short, it is incoherent to think with a blank predicate.

The way of mention: your new term makes its way into your mental vocabulary enclosed in quotes. The mental sentences that first connect the term to the pre-existing vocabulary are metalinguistic—they are sentences about the term or predicate itself, not about the category or property that it will come to represent. No problems here: it is perfectly coherent to think *about* a meaningless predicate. What metalinguistic thoughts, then, might serve to connect the new predicate to preexisting terms? Thoughts, presumably, that explicitly endow the term with semantic properties, that is, thoughts that declare the term to have a certain definition or a certain extension (or a certain conceptual role or . . .). But *now* there is a problem: such thoughts

will create analytic truths involving the new concept, yet it seems that there are no such truths.

You can see why the standard view of concept acquisition is compelling: the rival view of acquisition is logically unworkable. Yet the standard view predicts the existence of analyticities that are not in fact there. It is empirically unworkable.

Is there a third way? The dilemma I have posed runs almost exactly parallel to a dilemma posed by Fodor (1981) for what psychologists call supervised concept learning, that is, learning the preexisting concepts of other people, as in language acquisition. On the standard view of supervised concept learning, what the learner is trying to learn is a definition for the new concept; to learn the word 'dog', for example, they formulate various hypotheses about the correct definition of dog—where the correct definition is simply the definition employed by the local linguistic community—and then test their hypotheses against observed patterns of use. Fodor presents evidence that very few words have definitions, and argues that, since the existence of definitions is presupposed by the standard view, the standard view must be mistaken.

If supervised concept learners are not learning definitions, then what are they learning, asks Fodor? For more or less the reason given above—the logical incoherence of the alternatives—Fodor concludes that they are not learning at all. Rather, all concepts are innate.

Innateness need not mean out-and-out preformationism. As Margolis (1998) and Laurence and Margolis (2002) have observed (following earlier suggestions by Samet, Sterelny, and others), Fodor's argument leaves room for alternative possibilities. It rules out acquisition by "learning" in the traditional and most familiar sense, in which learning involves hypothesis formation and testing, but not the prospect that concepts make their way into the head in some other, more underhand way. Margolis suggests that what is innate about natural kind concepts, in particular, might not be concepts of individual kinds but rather a mechanism for acquiring new natural kind concepts without

explicit hypothesis formation. This mechanism would, on sighting a new kind such as the schwanns, manufacture a blank predicate ('schwann') and then "link" that predicate, or "put it in association with" the observed properties of the kind. Although Margolis does not say so, you might suppose that the linking process is automatic, unconscious, perhaps sub-personal.[13]

Such a suggestion is, for my purposes here, suggestive but critically incomplete. What is the nature of the "link" or "association"? Does it, when it is first established, use or mention the new predicate? If the former, how is it coherent? If the latter, does it create analytic truths? Let me flesh out Margolis's proposal in such a way as to answer these questions and to provide an account of the acquisition of minimalist natural kind concepts that entails no analyticities.[14]

Back to the misty morning schwann encounter. You have just come across these novel blue-feathered, purple-beaked, tromboning, non-swimmers. You, or your mind, is moved to create a new concept to capture what you take to be an as-yet unknown natural kind. Off the mental press rolls a suitable new predicate: 'schwann'. All that remains is to connect it, in the minimalist's causal way, to the putative characteristic properties of the new kind: blue feathers, purple beak, and so on. How is it done?

Let me propose a "Margolis module" (using the term 'module' loosely rather than in Fodor's (1983) strict sense). This device in the head does

13. Weiskopf (2008) also suggests that there are kinds of learning, or at least kinds of learning-like acquisition, other than the hypothesis formation and testing that Fodor's argument assumes. He proposes a model for concept acquisition of the descriptivist Kripkean reference-fixing variety: to introduce a concept, the learner picks out some feature of the world using a description, and then declares that feature to be the extension of a new blank predicate (which takes on the usual Kripkean semantic properties: direct reference, rigidity, and so on). Like all varieties of descriptivist reference-fixing, this looks to create analyticities, so it is of no help to me here.

14. What I will present is not strictly an extension of Margolis's proposal, since Margolis assumes a psychological essentialist theory of natural kind concepts that is inconsistent with my causal minimalism. But as far as I can see, Margolis's essentialism does not play an ineliminable role in his theory of acquisition.

what you, the learner, rationally cannot do: it takes the blank predicate 'schwann' and manufactures new mental sentences of the form *Something about schwanns causes P*, for each characteristic property *P*—thus, *Something about schwanns causes the growth of blue feathers*, *Something about schwanns causes the growth and use of a tromboning facility*, and so on—and then it drops these sentences into your belief box. (Alternatively, if you are a probabilistic cognizer, it drops them into your mental hypothesis space, with subjective probabilities attached.)

Without exercising any rational control over the process, then, you find yourself with some new "beliefs". I use the scare quotes because the sentences that have appeared in the belief box contain a heretofore meaningless predicate, 'schwann'. How, then, can the sentences be beliefs? I propose that the appearance in the belief box itself gives the predicate cognitive significance, and by way of that cognitive significance all the semantic properties that it needs. The new "beliefs" are in fact beliefs, and so a new minimalist concept of schwanns is successfully installed in the head.

How does the process implemented by the Margolis module give 'schwann' cognitive and semantic significance? Cognitive significance comes from the new term's being embedded in sentences that, in virtue of their form, acquire upon insertion into the belief box a certain inferential role. If you have sentences in the belief box of the form *Something about schwanns causes blue feathers*, *Something about schwanns causes tromboning*, and so on, then you will infer that an organism with the corresponding appearances and behaviors is a schwann. Conversely, if you believe that something is a schwann, you can infer using these same sentences that it will have the characteristic schwann appearances and behaviors, that is, the behaviors linked to schwannhood by the causal hypotheses.[15]

Further you can adjust your beliefs about schwanns as new information

15. These inferences are all, of course, defeasible. (Strevens 2000) describes some typical defeaters.

comes to light, either adding to the original stock of causal hypotheses as new properties are discovered or amending them if they turn out to be inaccurate—if it turns out, for example, that only the males of the species, or the local representatives, are purple-beaked.[16] Thus, if you treat 'schwann' as a perfectly well-defined, meaningful mental predicate and reason accordingly, it will rise to the occasion. You will make just the kind of reasonable-looking, useful inferences that you would make if you had a genuine schwann concept.

And yet—is it really a concept of schwanns? For that, it would have to have as its extension this particular species of blue-feathered, purple-beaked, tromboning landlubbers. The Margolis module makes no metalinguistic declarations—it makes no declarations at all—so if its activities are to give an extension to 'schwann', the extension must come along implicitly with the term's cognitive role. Margolis himself (and later, Margolis and Laurence) suggest for this purpose a version of a causal covariance theory of reference, on which a term refers to just those things that token the concept in the right sort of way. (Such theories differ in what they consider to be the "right sort of way"; Margolis and Laurence tentatively favor Fodor's own asymmetric dependence theory.) A closely related theory counts as a term's reference whatever the term's user is currently disposed, in ideal conditions, to predicate it of (Wilson 1982). I myself prefer a more forward-looking dispositional theory of reference advocated by Boyd (1988, 195):

> Roughly, and for nondegenerate cases, a term *t* refers to a kind . . . *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k*.

I will not try to defend such a view here, however; it is I hope sufficient to observe that there are a number of conceptions of reference on which the

16. In this case what happens is: you find organisms that, on the basis of properties other than beak color, you infer to be schwanns, that appear to be healthy, but whose beaks are not purple.

Margolis module can give an extension to a new natural kind concept without doing so explicitly. (You might alternatively, in a more Wittgensteinian frame of mind, claim that cognitive significance is enough and try to do without reference altogether.)

The Margolis module gives 'schwann' cognitive significance and an extension, then. (Or more exactly, the potential to have an extension; on a sophisticated theory of reference such as the forward-looking Boydean dispositional account, it is possible for a term to fail to refer even if it is actively traded on the cognitive exchange.) How does 'schwann' get its other semantic properties?

It does not get any other semantic properties. Or at least, if it gets some, the Margolis module is not responsible. The module does not give 'schwann' a definition. It does not give it a Fregean sense or a mode of presentation. It does not give it a "primary" or "epistemic" intension in Chalmers's sense. It does give the term a conceptual role—as must any process that introduces a substantial concept into the mental inventory—but it does not attribute semantic significance to that role, which is to say, the term is not somehow semantically linked to that role for the duration of its existence.

Because the Margolis module creates no semantic facts apart from facts about reference, and reference is fixed without any use of descriptions, the Margolis module creates no new analytic truths.

Let me develop this claim by dealing with two separate worries. First, might it not be that, contrary to what was claimed immediately above, the Margolis module does single out a certain conceptual role for 'schwann' as semantically special, as belonging to 'schwann' as of semantic right?

To see that this is not so, note that every one of the hypotheses by which the Margolis module bootstraps 'schwann' into cognitive significance might be judged, in the light of later evidence, to be false. We might, for example, discover that what we took to be characteristic of the schwann species is only one way that the species manifests itself. Some schwanns are blue-

feathered, purple-beaked, and have a trombone call, but others are pink-feathered, yellow-beaked, and sound more like a French horn. We might then further discover that it is the pink-feathered, yellow-beaked horn-players that are the norm; the schwanns that prompted our Margolis module's original act of concept creation acquired their appearances only because of powerful but ephemeral industrial pollutants in the environs. The sentences that the module deposited in the belief box when we first sighted the mutant schwanns were sufficient to get this train of reasoning going; their historical role however in no way stands in the way of their being later entirely abandoned. More generally, the sentences that the Margolis module places into the belief box arrive with no strings attached. They function therefore as purely empirical hypotheses, changing freely in response to the empirical evidence.

The second worry: won't any act of reference-fixing create analytic truths? Take, for example, a simple dispositional theory of reference on which a concept refers to whatever things its possessor would apply it to, under ideal conditions. Consider some natural kind concept K; let B be the complete set of beliefs in which that concept figures (for the concept's possessor). Plausibly, B completely determines what will be counted as falling under K in ideal conditions. In that case, there is some function from belief sets such as B to descriptions that pick out whatever someone with that belief set would count, under ideal conditions, as falling under K. Call this function $\phi(\cdot)$. Is it not an analytic truth for a thinker with belief set B that all Ks are $\phi(B)$?

Inasmuch as it is true entirely in virtue of semantic facts, yes. Because, however, the objectionable feature of analytic truths is not their truth in virtue of meaning per se, but rather their immunity to empirical disconfirmation, the creation of an analytic truth is a problem only if the truth is knowable a priori. There are two reasons to think that *Ks are* $\phi(B)$ is not known a priori by K's possessor. First, some aspects of $\phi(\cdot)$ presumably depend on facts about the possessor's psychological makeup, which may be knowable only a posteriori. Second, the correctness of the dispositional theory of reference,

and so the nature of $\phi(\cdot)$, is not known a priori if 'reference' is a theoretical term concerning the extension of which, as for other theoretical terms, there are no a priori truths.

A new topic: how does the Margolis module solve the use/mention dilemma? Is what the module does "mention"? Hardly. The process makes no metalinguistic assertions; the representations that it puts in the belief box do not mention the term 'schwann'.

Is it "use", then? Not exactly. As I characterized it above, the "use" mode of acquisition involves taking a reasoned attitude toward hypotheses using the term 'schwann'. You might interpret this rather loose description as follows: what characterizes "use" acquisition is that hypotheses using the new term, such as *Something about schwanns causes blue feathers*, gain their initial entry to the belief box by a process of reasoning. How to reason, though, about a hypothesis containing a manifestly empty term?

The Margolis module sidesteps this problem by getting the sentence into the belief box non-inferentially; it deploys the empty term in a position that implies imminent use, but then avoids using it.

Can we conceive of an empirical belief making its way into the belief box non-inferentially? Absolutely. Let me give you a very familiar example: perception. When I look at a table, according to most modern theories of perception, my mind delivers into my belief box, without any inference on my part about appearances or the reliability of vision, a mental sentence: "There is a table in front of me".[17] The belief simply appears. I can remove it if it seems suspect, but I cannot prevent it appearing in the first place. The sentence's entry into the belief box is automatic, unconscious, non-inferential.

The process by which the Margolis module puts sentences into the belief box has more or less the same character, I propose. It is non-inferential, but

17. That formulation is a little crude, of course, since the table belief is not always an occurrent thought, but then the belief box itself is a rather crude metaphor. Further, there is presumably something inference-like going on at the sub-personal level; the remarks in this paragraph concern the intentional level.

that does not mean that it has no place in intentional psychology—any more than perception should be excluded from intentional psychology. Unlike perception, however, it appears to have no place in our folk psychology. Learning, inference, stipulation, definition, perception: we are familiar with all of these. But we are not familiar with, and have no name for, the sort of process enacted by the Margolis module.

Let me give the process a name: *introjection*. In introjection, then, a new term is introduced into the mental vocabulary by way of something like the following three steps:

1. A new, blank mental predicate is manufactured.

2. The predicate is inserted into one or more sentences containing pre-existing mental vocabulary.

3. The resulting sentences are placed non-inferentially into the belief box (or into the partial belief box, with subjective probabilities attached).

The sentences in question might be deeply theoretical or they might be rather superficial: while the schwann concept and other natural kind concepts are introduced by causal hypotheses, the concept of, say, Santa Claus might be introduced by sentences such as *Santa wears red and has a big white beard*, *Santa brings presents at Christmas*, and so on.

Whether the sentences introduced by introjection are shallow or meaty, when things go well they will give the novel term so introduced a sufficiently rich inferential role that, in the relevant epistemic context, categorization and projection will be possible. Empirical inquiry—inductive logic bringing evidence to bear on the newly introjected hypotheses, or bringing the hypotheses to bear on other hypotheses—takes over from there. There is no guarantee, of course, that introjection will succeed in creating a concept with a determinate reference or a useful conceptual role, but then any theory of concept acquisition should, if it is to do justice to psychological fact, make space for half-baked, "ill-defined", or ultimately defective concepts.

Note that introjection, although motivated here by causal minimalism, is not proprietary to that view. A prototype concept could be acquired by introjection, if sentences asserting a statistical connection between category membership and characteristic properties were introjected, as could an essentialist concept, if a sentence asserting a metaphysical connection between category membership and possession of an essence was introjected. (In this latter case, what ends up in the belief box is a metaphysical hypothesis, but one that is subject to empirical disconfirmation.)

What introjective processes are found in the mind? There is, first, the process for acquiring natural kind concepts implemented by the Margolis module. Perhaps other varieties of concepts—artifact concepts, for example—have their own proprietary introjective acquisition procedures.

A second procedure that might be interpreted as a kind of introjection is Carey's (2009) "Quinian bootstrapping", a concept acquisition process in which new concepts are given cognitive significance by being attached to a pre-existing but uninterpreted formal structure (uninterpreted, that is, by the learner). Carey argues for example that children acquire the concept of a natural number in this way, by attaching preexisting "proto-number" concepts (which exist only for the lower single-digit numbers) to the grammatical structure of their language's system for enumerating the natural numbers, that is, for counting from one to any number you like. The mechanism at work here might be understood as follows: the axioms or other basic statements of the formal linguistic structure in question are copied non-inferentially into the belief box, partially filled out by pre-existing proto-number concepts. That would be a kind of introjection.

What other mental processes are introjective? Perhaps conceptual innovation in science?

5. Theoretical Introjection

Introjection introduces novel terms to a thinker's mental vocabulary without creating analytic truths. In science, novel theoretical terms are typically introduced to scientific vocabulary without creating analytic truths—a mystery. Could it be some sort of introjection that creates new scientific terms?

What would this "theoretical introjection" look like? When a new term such as 'fitness' is introduced, it would make its first appearance as a part of one or more empirical hypotheses with no analytic corollaries. In psychological terms, a new mental predicate 'fitness' would be minted,[18] and sentences such as *Fitness is proportional to expected number of offspring*, *The fitter of two competing variants tends to replace the other*, and *Physiologically identical organisms are equally fit* would then appear in the belief box, where they would function as empirical hypotheses, guiding inferences about fitness but also themselves subject to refutation or confirmation by the facts.

To defend the introjective view of the introduction of 'fitness', it would be necessary to show how these original, introjected, empirical hypotheses provide an inductive ground sufficiently firm to support all of the development of evolutionary biology that has unfolded since—though perhaps it is enough to note that, since no one can point to a widely acknowledged analytic truth about fitness, the development of evolutionary biology must have been built on statements about fitness that lack the apodictic authority of analytic truths, even if we cannot say exactly how. In any case, to give a full defense of the hypothesis that the acquisition of new scientific concepts proceeds by introjection is not, for want of space, my goal here. Let me rather ask a more limited question: how is the introjection of scientific concepts possible? By

18. The English word 'fitness' was not of course in Darwin's time a blank predicate, but the mental term to which the biological sense of the word came to be attached was, I surmise, distinct from the pre-Darwinian English word's opposite number in Mentalese. That is, when evolutionary biology introduced the notion of fitness it reused and repurposed a linguistic token, but the mind manufactured an entirely novel mental token—so I speculate.

what psychological mechanism might sentences such as those concerning fitness above make their way non-inferentially into the head, in the absence of a definition of 'fitness' or other explicit reference-fixing act?

Inspired by the Margolis module, you might posit some device in the head that manufactures new scientific concepts. But this seems unlikely. The Margolis box is plausible only because it operates within certain tight constraints: it responds only to (apparently) new chemical substances and biological kinds; it seizes only on certain properties of those kinds, that is, certain sorts of appearances and behaviors that are considered to be chemically or biologically relevant; and it builds new hypotheses with a quite specific form: *Something about* ____*s causes* ____. A box that responds to the full gamut of science's conceptual needs strikes us as a fantasy.

Can new scientific concepts be acquired through Carey's Quinian bootstrapping? In some cases, perhaps: an uninterpreted mathematical model, built to save a certain phenomenon, might be copied to the belief box (or hypothesis box) with blank predicates attached to some of its internal components; these predicates would then function as new concepts given cognitive significance, but not definitions, by their role in the model. Surely not every example of conceptual innovation can be explained in this way, however; only in some cases does an uninterpreted formal model predate a new concept's introduction. Perhaps more important, what is missing here is a specification of the conditions under which the copying operation is carried out.

Let me propose an account of theoretical introjection according to which certain kinds of linguistic acts or their mental equivalents, which are in their overt form not unlike stipulations, bring about not definition but introjection in the minds of both speakers and listeners.

Consider, as an example, the way in which the notion of general intelligence might have been introduced:[19]

19. The canonical introduction of this term in Spearman (1904) is rather more diffuse, but certainly not at odds with my suggestion here.

> Maybe there is a single capacity, 'general intelligence', that causes the observed correlation between children's levels of performance in a wide range of visual, verbal, and mathematical tests.

Such a neologistic act might be interpreted as a definition: let 'general intelligence' mean "the capacity that causes the observed correlation etc.". Or it might be interpreted as an act of descriptivist reference-fixing: let 'general intelligence' be whatever best satisfies (and satisfies well enough) the description "the capacity that causes the observed correlation etc.".

I suggest that, insofar as these interpretations are supposed to capture what goes on in the mind of either the speaker or the listener when the words above are uttered, they are mistaken. Certainly, we could have been psychologically constituted so as to understand words of this form as defining or reference-fixing; in that case, our reaction to the words would be to produce a new mental predicate and, following the way of "mention", to give it semantic significance by explicitly attaching to it the definition or the extension in question. In other words, we would make a mental declaration of the form: *I shall use 'general intelligence' to mean, or to refer to, ...*

But this is not our reaction. Rather, we respond by performing an act of introjection. That is to say, we manufacture a new mental predicate, which I will somewhat misleadingly write 'general intelligence'; we use that predicate and pre-existing mental vocabulary to build a mental sentence of the form *General intelligence causes the observed correlation between children's levels of performance in a wide range of visual, verbal, and mathematical tests*; and we place this sentence non-inferentially into our belief boxes (or better, into our partial belief boxes, with an appropriate subjective probability attached). Then we begin to reason inductively with the new sentence, conducting statistical tests, for example, to determine the degree of correlation between abilities that can be ascribed to general intelligence, and perhaps eventually—if the evidence pushes us in this direction—concluding that there is no such thing

as general intelligence.[20]

Why is general intelligence in quotes, suggesting mention rather than use, in the hypothetical act of introduction above? Because the linguistic term 'general intelligence' is introduced by definition: it is defined as equivalent to the unnamed mental term introduced by introjection. This will be clearer if the introductory act is rewritten as two sentences:

> Maybe there is a single capacity that causes the observed correlation between children's levels of performance in a wide range of visual, verbal, and mathematical tests. Call it 'general intelligence'.

On my view, the first sentence introduces a novel concept by introjection. The second sentence is exactly what it appears to be: an act that explicitly confers semantic significance on the linguistic item 'general intelligence' by attaching it to the newly created concept. This analysis suggests that merely to posit an unobservable cause of some observed correlation is to introduce a new concept by introjection, a concept that is bound to the descriptive phrase "unobserved cause of such and such a correlation" neither by definition nor by reference-fixing, but rather by its being posited as such a cause in an empirical hypothesis.

Positing common causes, or more generally hidden explainers, triggers concept creation by introjection, then. One way that such explainers are suggested is analogy (Nersessian 2010). Take the case of the caloric fluid theory of heat, for example. Before the theory is formulated, there exist a set of hypotheses about ordinary fluids. They flow downhill, in every available direction, in a continuous wave, traveling faster as the slope is steeper. The notion of caloric fluid might be introduced as follows:

20. Observe that a thing's non-existence can always be inferred by a purely inductive route. If we believe that *g*, if it exists, is the cause of such and such a correlation, and there turns out to be no such correlation, it is reasonable to conclude, provisionally, that *g* does not exist. You do not need to *define g* as the putative cause of the correlation to reach this conclusion; it is enough to believe it.

> Perhaps heat is like a fluid. This 'caloric fluid' flows down temperature differentials, traveling faster as the differential is steeper, and so on.[21]

The pre-existing hypotheses serve here as templates (after they are amended so that temperature differential is substituted for slope). Such templates might be used for the explicit definition or reference-fixing of a new mental term corresponding to the linguistic term 'caloric fluid'. Might, but are not: I propose that rather, in the course of the analogy, the new mental term is simply dropped into the templates and the novel sentences so constructed pushed into the belief box. There, like all the contents of the belief box, they are treated as empirical hypotheses about a certain kind of substance. No definitions, no analyticities, are brought into existence.[22]

Introjection may also be triggered by the postulation of a hidden mechanism:

> I surmise that there is a process—call it 'introjection'—that under certain circumstances puts sentences containing blank predicates into the belief box non-inferentially.

This is not a definition, as further empirical inquiry might show that it is mistaken in certain ways. Perhaps, for example, it will turn out that the new predicates introduced by introjection are not always entirely blank—perhaps they carry with them some pre-theoretical connotations or inferential roles, as when scientists repurpose a pre-existing term such as 'fitness' or 'intelligence'.

21. Lavoisier's (1783) introduction of the term is motivated by the ability of caloric fluid to explain thermal expansion. He writes "One can hardly conceive of such phenomena [expansion and contraction] without admitting the existence of a certain fluid, the accumulation of which is the cause of heat and the absence of which is the cause of cold. Undoubtedly this fluid insinuates itself between the particles of bodies, pushing them apart and occupying the spaces that they leave between them. I call . . . this fluid, whatever it is, caloric fluid (*fluide igné*) . . . "

22. The use of pre-existing templates makes this process closer to Carey's Quinian bootstrapping; perhaps it is a mistake to try to draw a bright line between the two.

Or perhaps introjection is not wholly non-inferential—perhaps, for example, when an introjected hypothesis is placed in the partial belief box, its subjective probability is determined in part by the credibility of the scientist making the introjection-inducing posit. I am not endorsing these suggestions, but they do seem to me to be open possibilities.

My proposed approach to theoretical introjection—that certain kinds of theoretical posits induce introjection in both speaker and listeners, creating the concepts needed to make the posits complete—requires a psychology. The psychology will specify in principle and not merely by example what forms of speech or thought will prompt introjection, and what the introjected hypotheses will be. It will answer the question whether Quinian bootstrapping is a form of introjection. It will introduce a new form of intentional thought to cognitive psychology—and perhaps to folk psychology too. And it will explain how theoretical terms can be introduced into science without creating hypothesis-like analytic truths.

Acknowledgements

Thanks to the Corridor Reading Group, to audiences at Yale, Harvard, and NYU, and to Nick Shea, Susan Carey, and an anonymous referee for their valuable comments and suggestions.

References

Boyd, R. (1988). How to be a moral realist. In G. Sayre-McCord (ed.), *Essays on Moral Realism*. Cornell University Press, Ithaca, NY.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press, Oxford.

Carnap, R. (1937). *The Logical Syntax of Language*. Translated by A. S. C. von Zeppelin). Routledge, London.

Chalmers, D. J. (2006). The foundations of two-dimensional semantics. In M. García-Carpintero and J. Macià (eds.), *Two-Dimensional Semantics*. Oxford University Press, Oxford.

Donnellan, K. S. (1966). Reference and definite descriptions. *Philosophical Review* 75:281–304.

Enç, B. (1976). Reference of theoretical terms. *Noûs* 10:261–282.

Feyerabend, P. K. (1962). Explanation, reduction, and empiricism. In H. Feigl and G. Maxwell (eds.), *Scientific Explanation, Space, and Time*, volume 3 of *Minnesota Studies in the Philosophy of Science*. University of Minnesota Press, Minneapolis.

Fodor, J. A. (1981). The present status of the innateness controversy. In *Representations*, pp. 257–316. MIT Press, Cambridge, MA.

———. (1983). *The Modularity of Mind*. MIT Press, Cambridge, MA.

———. (2008). *LOT 2: The Language of Thought Revisited*. Oxford University Press, Oxford.

Gelman, S. A. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford University Press, Oxford.

Gillespie, J. H. (1977). Natural selection for variances in offspring numbers: A new evolutionary principle. *The American Naturalist* 111:1010–1014.

Hardin, C. and A. Rosenberg. (1982). In defense of convergent realism. *Philosophy of Science* 49:604–615.

Keil, F. C. (1989). *Concepts, Kinds and Conceptual Development*. MIT Press, Cambridge, MA.

Kitcher, P. (1978). Theories, theorists, and theoretical change. *Philosophical Review* 87:519–547.

Kripke, S. (1977). Speaker's reference and semantic reference. *Midwest Studies in Philosophy* 2:255–276.

———. (1980). *Naming and Necessity*. Harvard University Press, Cambridge, MA.

Kroon, F. W. (1987). Causal descriptivism. *Australasian Journal of Philosophy* 65:1–17.

Laurence, S. and E. Margolis. (2002). Radical concept nativism. *Cognition* 86:25–55.

Lavoisier, A. L. (1783). Réflexions sur le phlogistique. *Mémoires de l'Académie des Sciences* Année 1783:505–538.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy* 67:427–446.

Margolis, E. (1998). How to acquire a concept. *Mind and Language* 13:347–369.

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press, Cambridge, MA.

Nersessian, N. J. (2010). *Creating Scientific Concepts*. MIT Press, Cambridge, MA.

Psillos, S. (1999). *Scientific Realism: How Science Tracks Truth*. Routledge, London.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review* 60:20–43.

———. (1963). Carnap and logical truth. In P. A. Schilpp (ed.), *The Philosophy of Rudolf Carnap*, volume 11 of *Library of the Living Philosophers*. Open Court, Chicago.

Spearman, C. (1904). "General Intelligence" objectively determined and measured. *American Journal of Psychology,* 15:201–292.

Sterelny, K. (1983). Natural kind terms. *Pacific Philosophical Quarterly* 64:110–125.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition* 74:149–175.

———. (2008a). *Depth: An Account of Scientific Explanation*. Harvard University Press, Cambridge, MA.

———. (2008b). Physically contingent laws and counterfactual support. *Philosopher's Imprint* 8(8).

———. (Forthcoming). Ceteris paribus hedges: Causal voodoo that works. *Journal of Philosophy*.

Suppe, F. (1977). The search for philosophic understanding of scientific theories. In F. Suppe (ed.), *The Structure of Scientific Theories*. Second edition. University of Illinois Press, Urbana and Chicago.

Weiskopf, D. A. (2008). The origins of concepts. *Philosophical Studies* 140:359–384.

Wilson, M. (1982). Predicate meets property. *Philosophical Review* 91:549–589.