9/19/2006

Naturalism Week 3

Overall Plan:

1. Preliminary: There are ***two levels* of supervenience** (reducibility, token-token identity, etc.) **claims** that one can considered, and it is important to keep them distinct:

a) *Particular, ground-level* supervenience (etc.) claims, which say that, for particular values of 'φ' and 'ψ', ψ-vocbulary/facts supervene (etc.) on φ-vocabulary/facts. In this sense, one might be a scientific naturalist of the physicalist sort about, say, intentional vocabulary, but *not* about moral normative vocabulary (perhaps because one has an *error theory* [Mackie]about it, or one takes it not to be *descriptive* or *explanatory*, but to perform some different linguistic function).

b) *General, metaphysical* supervenience (etc.) claims, to the effect that *everything real* (*all* values of ψ) supervenes (etc.) on, say, the fundamental physical (φ). Sellars's *scientia mensura* is a claim of this sort: "In the dimension of describing and explaining, science is the measure of all things, of those that are, that they are, and of those that are not, that they are not." (from "Philosophy and the Scientific Image of Man").

Claims of the second, grander, sort must, it seems, be motivated by some form of the *trans-domain hegemony inference*:

i) *From* Intramural excellence: the exhibition by the favored φ of some kind of special (presumably *unique*) privilege within its own domain (paradigmatically, causal or explanatory *completeness* or *closedness*, a distinctive kind of explanatory *adequacy* or *success*);

ii) *To* Extramural Authority or Privilege: the possession by the favored φ of a *different* kind of special (again presumably unique—i.e. no *other* vocabulary has *this* sort of privilege) with respect to *other* domains, for instance that facts statable in other vocabularies must *supervene* on, or be *reducible* to the φ-facts.

The challenge for those making claims of this grander, metaphysical or ontological sort (and there is also a *methodological* or *epistemological* version that says that the *methods* of natural science have proven themselves so successful within their own domain that they deserve to be thought of not just as especially good or promising, but as the *only* methods for delivering genuine (empirical?) *knowledge*) is to specify a kind of *intramural excellence* in (i) that genuinely justifies the specific sort of claim to *extramural authority* in (ii).

Compare the political analog: Our system (that say, of the English in the days of the British Empire, 21st century U.S. Republicans, evangelical fundamentalist Muslims…) works so well at home (exhibits virtue of type (i)) that that system ought to govern (to possess authority of type (ii) over) other domains.

2. Recall the general situation with respect to the understanding of the *relations* between possible naturalistic base vocabularies (whether physicalistic, more broadly natural-scientific so as to include chemistry and biology understood as not necessarily reducible to physics, as including the special natural sciences such as geology and astronomy, or as even broader) and various possible target vocabularies of antecedent

interest (psychological in the sense of sapience-intentionality, or of sentience-consciousness, various sorts of normative vocabulary), whether thought of as underwriting particular, ground-level claims or the general, metaphysical conclusions of trans-domain hegemony inferences. The aim is to specify relations such that some naturalistic thesis will turn out to be both arguably *true* and evidently *interesting*. Recent discussion has described a pendulum-like swing between various sorts of *reductionism*, which seem not likely to be *true*, and various forms of *supervenience*, that don't seem to be sufficiently *interesting*.

a) From roughly 1965 to 1980, Anglophone philosophers gave up on *reducibility*, because of Putnam and Fodor's arguments concerning:
   i) Many-levels (e.g. Putnam's "round peg in a square hole" argument); and
   ii) Multiple realizability (of functional properties by physically specifiable objects).

b) Since about 1980, awareness has been growing that at least some forms of *supervenience* don't give us enough naturalism to be worth having. Those arguments are summarized by Horgan (see (6) below).

3. Recap **Beth definability**, which raises questions about the viability of the *distinction* between global supervenience and Carnap-Nagel reducibility. It really only addresses the *definability* portion of the Carnap-Nagel *definability* (of terms) & *derivability* (of laws) notion. So if it were taken at face value, we would have an intermediate position: definability without derivability. (As noted below, this is also where we get if we combine an argument of Kim's with Stalnaker's argument laid out in his appendix. See (8) below.) This is more than anomalous monism or mere token-token identity, since it is type identity (definability) without any way of mapping the (modally robust or counterfactual-supporting) *inferences* of the ψ-theory onto those of the φ-theory. Note that on even a *weakly* inferentialist account of conceptual content, this is *incoherent*, since the concepts are individuated at least in part by the (multipremise, modally robust or counterfactual-supporting) inferences they are essentially involved in as premises and conclusions.

a) (Modally strong) global supervenience of ψ-vocabulary (facts) on ϕ-vocabulary (facts) holds in case no two *possible* worlds can agree on all the facts statable in the ϕ-vocabulary (ϕ-facts) and disagree on the facts statable in the ψ-vocabulary (the ψ-facts).

b) Q: But what sense of '*possible* world' is wanted? (Stalnaker is good on this issue.)

c) It can't be <u>*logically* possible</u> because it is not in general *logically* inconsistent to have ψ-differences without ϕ-differences.

d) It might be *metaphysically* possible worlds—but what does that mean in this context? The Kripkean sense, which seems to derive from *semantics* does not seem helpful here.

e) And if we want *physically* (ϕ-vocab/facts) possible worlds, then we seem to be going around in a circle (if naturalism is our question), since then what we want to know is whether it is *physically* possible for the ψ-facts not to supervene on the *physical* (ϕ-)facts.

f) The antecedent condition of the Beth theorem restricts us, not to a set of *possible worlds*, but to a set of *models*, that is, *algebraic structures*, consisting of a *domain* and a set of *relations*. Q: *Which* set of relations? A: All those that *satisfy*, i.e. are *models* of, the (hypothetical) first-order *theory* that relates the ϕ-vocabulary/facts to the ψ-vocabulary/facts.

g) Recall that Carnap starts to back away from full (Carnap-Nagel) reduction by invoking *partial reduction formulae*, which are not biconditionals, but merely *some* set of conditionals relating ϕ- and ψ-vocabulary.

h) Beth considers what it would be to have *enough* such "partial reduction formulae". What is 'enough'? Intuitively, enough to "fix the meanings" of the ψ-vocabulary in terms of the ϕ-vocabulary.

i) And his way of settling *that* is *semantic*, in the model-theoretic sense: to specify that there cannot be two *models* satisfying the *theory* (the "partial reduction formulae") that *agree* on which ϕ-sentences are satisfied by the model, but *dis*agree on which ψ-sentences they satisfy.

j) On that condition, which we could characterize as saying not that ψ-vocabulary/facts *supervene* on ϕ-vocabulary/facts, but that ψ-vocabulary/facts are *implicitly definable* in terms of ϕ-vocabulary/facts, Beth proves that it is possible to construct *explicit definitions* of all ψ-terms in the form of *biconditionals* relating them to ϕ-terms.

k) So the big question is how we should understand the difference between saying that:

i. There are no two X-possible (possible in sense X, which is, in effect, a *parameter* that we can fill in in different ways) *worlds* that agree on all the ϕ-sentences that are *true* in them, but disagree on which of the ψ-sentences are *true* (possible-world *supervenience*), on the one hand, and

ii. There are no two *models* of any first-order *theory* relating ϕ- and ψ-vocabulary/facts that *satisfy* all the same ϕ-sentences, and differ on which ψ-sentences are *satisfied* (model-theoretic *implicit definability*).

l) Triangulating this issue with two other ideas that may be instances of a common phenomenon:

i. I suggested that this question may be a way into thinking about a *deep* issue concerning the relation between modal *possible worlds* talk and algebraic *model-theoretic* talk in thinking about meanings (vocabulary) and metaphysics (facts).

ii. And that it is possible that this is *another* way into (what might turn out to be) the *same* issue that John Etchemendy pursues in his *The Concept of Logical Consequence*, which explores tensions between:

α) The Tarskian model-theoretic account of logical consequence—according to which S|=T iff every model that *satisfies* S also *satisfies* T; and

β) The *justification* of this formal model of logical consequence, which understands it as explicating the more-or-less intuitive sense we have (something we would like to be *entitled* to say) that S logically entails T iff it is *impossible* for S to be *true* and T *not* to be *true*, i.e. there is no (X-) *possible* world in which S is true and T is *not* true.

iii. We'll see that the same issue comes up again for the (model-theoretic, but couched in the language of possible worlds) proof that Stalnaker offers in his Appendix, showing a sense in which global supervenience entails local supervenience.

4. **Craig elimination**. This result addresses not just the *relation* between φ- and ψ-vocabularies/facts, but also the claim of the base φ-vocabulary to explanatory *sovereignty* or *completeness* within its own domain, which is the basis of its claim to hegemony of some kind over *other* vocabularies and domains. For, if taken at face value, it would show that *every* vocabulary has just the *same* sort of explanatory sovereignty over or completeness within *its* own domain.

Craig's Elimination Theorem offers another example of a result, syntactic this time, concerning semantic relations among vocabularies. Two lessons: its generality and (hence) triviality; and the subtle distinction of kinds of vocabulary it enforces.

a) Statement: Given any language L, any recursively axiomatizable first-order quantificational theory T on L, and any partition of the sentences of L into a recursively specifiable subset L' of L, on the one hand, and L-L', on the other, there is a recursive axiomatization of T'—(=T/L') the fragment of T that consists solely of sentences of L'—that itself consists solely of sentences of L'.

b) Hempel (in his classic 1958 essay "The Theoretician's Dilemma") takes L' to be the *observational* fragment of a scientific language: that is, to consist of all sentences couched *entirely* in *observational vocabulary*. So we have the observational vocabulary V_O, and L_O as the set of sentences in which the *only* non-logical vocabulary is from V_O. Then the theorem guarantees that for *any* theory T, there is a recursive axiomatization that

i. Has as its consequences exactly the purely *observational* consequences of T and
ii. Is itself couched entirely in observational vocabulary.

Hempel sees this as offering a challenge to our understanding of the utility of *theoretical* vocabulary. If what we care about is codifying possible observations, then the theorem seems to say that theory is *superfluous*. (Hence the name: "Craig *Elimination* Theorem.") We can always get the same effect while remaining wholly within the observational vocabulary. Of course, there may be other virtues this formulation lacks: simplicity (it will, for instance, under very general circumstances *always* have an *infinite* number of axioms, even if T had only a finite number), support of inductions, explanatory insight. (We'll see just how true this is, and why, when we look at how the construction works.)

c) The first thing I want to point out is how *general* this theorem is. It is proved entirely at the level of *sentences*, so we can pick out our privileged subset of the language by looking not just at non-logical *predicates* drawn from some vocabulary, but also at *singular terms*, or even *sentential operators*. The *only* requirement is that the inner set of sentences L' (the one Hempel took to be *observation* sentences) is *recursively specifiable*. This condition is easily satisfied if the set of sentences in question consists of all and only those containing tokens of a certain lexical type, or consisting entirely of vocabulary drawn from some finite list of lexical types.

d) So for instance, it has as a consequence that any theory involving *modal* vocabulary can be axiomatized entirely in *non-modal* vocabulary so as to have exactly the same *non-modal* consequences.

e) And any theory involving *normative* vocabulary can be axiomatized entirely in *non-normative* vocabulary so as to have exactly the same *non-normative* consequences.
f) And any theory involving *indexical* vocabulary can be axiomatized entirely in *non-indexical* vocabulary so as to have exactly the same *non- indexical* onsequences.
g) If the result shows the *dispensability* in a precise sense, of *theoretical* vocabulary, it does so equally for *modal, normative, indexical, intentional* or indeed *any* sub-vocabulary. Are we to conclude that *all* of these are *dispensable*, because Craig-eliminable?
h) Perhaps worse, the *converses* of all of these eliminations *also* are possible: doing away with *non*-normative vocabulary while leaving intact the *normative* results, doing away with *non*-modal vocabulary, and so on.
i) Notice that this result shows the emptiness of one common way of thinking about the nature of the privilege accorded to *physicalistic* vocabulary by high-church, Unity-of-Science naturalism.

i. For it might be thought that one thing that is special about physicalistic vocabulary is that it is *sovereign* or *self-sufficient* within its own domain, in that every phenomenon described entirely in physicalistic terms can be explained entirely in physicalistic terms.
ii. By contrast, one might think, some phenomena specified in *aesthetic* terms (say, some facts about the uniformity of a color or the smoothness of a curve) can only be explained by appealing to facts described in *physicalistic* terms, as well as purely aesthetic ones.
iii. Craig's theorem shows that for at least one clear sense of 'explains'—the Deductive-Nomological notion of derivability from a recursively axiomatizable theory—physicalistic vocabulary has no privilege in this regard at all. For to *any* theory that has consequences couched in aesthetic vocabulary, there corresponds a recursively axiomatizable theory that has just the same consequences couched in aesthetic vocabulary, and which is itself entirely formulated in aesthetic vocabulary.
iv. That is, at least this feature that naturalists might be tempted to appeal to in justifying the privilege of physicalistic (or other naturalistic) vocabulary—its explanatory self-sufficiency—turns out to be true of *every* vocabulary. So, for instance, in this same sense, *non-normative* vocabulary is eliminable: the normative shows up as an explanatorily self-sufficient realm in exactly the same sense that the natural is.

j) Austin famously said that any philosophical story consists of "the bit where you say it, and the bit where you take it back." Having said a bunch of *true* things about Craig's result, here comes the bit where we take it back, by seeing how unlikely it is that these claims will end up being *interesting*. (Though that does not at all mean we can't learn anything from them.)
k) What is going on here? When we look at how Craig's construction actually works, I think we'll see that the Craig eliminability of a vocabulary is *much* less significant than it might at first appear to be—as we should expect once we notice that *every* vocabulary is eliminable in this sense.

i. I said above that the recursive axiomatization Craig constructs will always contain an infinite number of axioms. What makes it recursive is that there is an algorithm for determining in a finite number of steps whether *or not* any given sentence is an axiom in that set.

ii. For Craig in effect takes every consequence of the original theory that is couched in the privileged vocabulary to be an axiom.

iii. There is no guarantee that that set will be recursively specifiable, so something else is needed. Craig uses a trick (indeed, his result is usually referred to by mathematicians as "Craig's trick" rather than "Craig's Theorem") that is at once clever and stupid. For any sentence p that is a consequence of T couched entirely in the privileged vocabulary (say, observational vocabulary), Craig uses as his corresponding axiom not p, but the conjunction of p with itself n times: p&p&p&p…&p. This is obviously logically equivalent to p. His trick consists in his choice of n. Since p is by hypothesis a consequence of the recursively specifiable theory T, there is a proof of it. Index those proofs numerically, for instance by assigning each such proof to its Gödel number. Then use the ordinal value in the well-ordering that results from the Gödel numbering. The constructed axioms in the privileged vocabulary can be ordered according to how many conjunctions they consist in. Then to check whether some sentence x in the privileged vocabulary is an axiom of the reconstructed theory T', one must just see whether it is a self-conjunction of length n, and if so, whether n is the Gödel number of a proof of the self-conjoined proposition in T. (So you need to use T in order to check—you can't dispense with it in favor of the new theory, T'). Both of those are recursive procedures.

<u>How the Proof of Craig's Elimination Theorem Works</u>:

The basic steps:

l) Want to take all the T/V_I formulae as axioms.

m) But need to be able to specify them recursively.

n) So take each of them as an n-fold conjunction of itself, with the n chosen cleverly to make it possible to determine recursively whether any n-fold conjunction is the conjunction of an 'axiom'.

o) Do that by appealing to the proof *in the full theory T*, of that sentence.

p) Use the Gödel number of the proof to determine n.

q) When explaining the basic steps, perhaps tell the story of my Sheffer operator, which uses the second argument-place to code the variable of quantification, in case the first one is an open formula. It does that by making it the conjunction of n formulae, for the n^{th} variable.

r) More careful statement of Proof. Preliminaries. We start with then notion of the <u>recursive axiomatizability</u> of a theory. It is a notion intermediate between <u>finite axiomatizability</u> and <u>recursively enumerable axiomatizability</u>. Unless one means one of these, it means nothing to say that a theory is 'axiomatizable': one can simply take all the sentence of the theory as axioms. Then all the sentences of the theory trivially follow logically from the axioms. The question is whether one can find some subset $A \subset T$ such that the deductive closure of A is T (i.e. A|—T) and:

i. A consists of only a finite number of sentences (finite axiomatizability);

ii. There is an algorithm that will, for any sentence in the language, determine in a finite number of steps whether or not it is an element of A (recursive axiomatizability);

iii. There is an algorithm that will, for any sentence in the language, in a finite number of steps settle that that sentence is an element of A if it is, but may not settle in a finite number of steps that it is *not* an element of A, if it is not (recursively enumerable axiomatizability).

s) Start with two vocabularies, a *complete* vocabulary V_C and an *inner* vocabulary V_I, thought of as sets of *sentences*, with $V_I \subseteq V_C$. A *recursive* theory T, formulated in V_C.

A theory $T \subseteq V_C$, is:

i) *Unrestricted* if it is *any* arbitrary subset of V_C;

ii) *Finitely axiomatizable* if there is a *finite* subset $A_T \subseteq T$ such that $A_T \vdash\!\!\!-T$.

iii) *Recursively axiomatizable* if there is a subset $A_T \subseteq T$ such that $A_T \vdash\!\!\!-T$, and there is an algorithm that for any sentence $s \varepsilon V_C$ will determine in a finite number of steps *whether or not* $s \varepsilon A_T$—but, if *not*, it may simply not give an answer.

iv) *Recursively enumerably axiomatizable* if there is a subset $A_T \subseteq T$ such that $A_T \vdash\!\!\!-T$, and there is an algorithm that for any sentence $s \varepsilon V_C$, *if* $s \varepsilon A_T$ will determine in a finite number of steps *that* $s \varepsilon A_T$ —but, if *not*, it may simply not give an answer.

t) Theorem: The theorem then states that if V_I is itself recursively specifiable, then there is a *recursive* set of axioms $A_{T'} \subseteq V_I$ such that $A_{T'} \vdash\!\!\!-T/V_I$.

u) So Craig's theorem shows that for *any* recursively specifiable inner sub-vocabulary, *any* theory that 'explains' all some set of facts statable in the *complete* vocabulary has a *recursive sub*-theory statable entirely in the *inner* vocabulary that entails *all* the consequences of the original theory that are statable entirely in the *inner* vocabulary. In that sense, *for any* recursive theory that explains some facts in the complete vocabulary—in the sense that they can be *derived* from it—*no matter what* the inner vocabulary is, it is *explanatorily complete* in the sense that all the facts *statable in that vocabulary* can be derived from a *recursive* theory *statable entirely in that inner vocabulary*.

v) Proof:

i. Consider all the sentences s_i that belong to the restriction T/V_I of T to the *inner* vocabulary. These are the *target* sentences, which must be *entailed* by the theory we are going to construct, which must *also* consist entirely of sentences of V_I.

ii. Since each of these $s_i \varepsilon V_I$ is a sentence of T, which is recursively axiomatizable by $A_T \subseteq T$, there is a *proof* of s_i from A_T. Call this proof $A_T(s_i)$.

iii. It consists of a finite set of sentences of V_I. So we can take the *conjunction* of those sentences, in the order in which they form the proof. Call this conjunction $C(A_T(s_i))$, which will be a sentence of V_I.

iv. We can now form the unique *Gödel number* of this sentence of V_I. [Explain briefly what this means and how it works.]

v. Since there are a finite number of s_i, and their Gödel numbers are unique, we can arrange them in a well-ordered linear sequence, from that with the smallest Gödel number to that with the largest. Call the position of each s_i in this sequence $n(C(A_T(s_i)))$.

vi. For each s_i, form $\&_n(s_i)$, which is $s_i \& s_i \& \ldots s_i$, $n(C(A_T(s_i)))$ times.

vii. Obviously, each $\&_n(s_i)$ is largely equivalent to s_i.

viii. Also obviously, $\{s_i: s_I \varepsilon T/V_I\} \vdash\!\!\dashv T/V_I$. So the set of all $\&_n(s_i)$ is an *axiomatization* of T/V_I. But there is no guarantee that it will be a *finite* set.

ix. We can, however, specify $\{\&_n(s_i): s_I \varepsilon T/V_I\}$ *recursively*. For to test for any $s \varepsilon V_I$ whether $s \varepsilon \{s_i: s_I \varepsilon T/V_I\}$ we just need to check how many identical elements of V_I s consists of. (It will always consist of at least 1 such element—namely itself—conjoined with itself, and it is guaranteed to be in V_I.) Call this number $n_\&(s)$, and the conjuncts (which may or may not be identical to s) $s_{n\&}$. Then $s \varepsilon \{\&_n(s_i): s_i \varepsilon T/V_I\}$ iff $n_\&(s)$ is the ordinal value of the Gödel number of a proof in T of s_i, that is, iff $n_\&(s) = n(C(A_T(s_i)))$.

w) The requirement that the partition of L be *recursively specifiable* actually does put a significant restriction on the vocabularies to which the theorem applies. List argues that it rules out the very application Hempel seizes on: observational vocabulary. For if the vocabulary in question is a vocabulary only in the *broad* sense, and not just the *narrow* sense that coincides with some lexically or syntactically identifiable repeatables, then the theorem does *not* apply. So, for instance, if "*demonstrative* or *deictic* vocabulary" refers not just to all tokenings of some demonstrative *types*, perhaps 'this' and 'that', but also to some tokenings of other types—for instance, 'the cat'—then there need be no recursive way of specifying sentences in which this vocabulary appears. In fact it is clear not only that other expressions *can* be used demonstratively, but also that in natural languages there are *no* lexical types *all* of whose tokenings are used demonstratively. 'this' and 'that' also have *anaphoric* uses. But note that this problem with, e.g., the idea of *observational* vocabulary is not a problem with *physicalistic* vocabulary, which arguably *is* recursively specifiable.

x) What should we conclude from the Craig Elimination Theorem? Primarily, I think, that we must be *very careful* in formulating the *explanatory completeness* or *closedness* condition on potential naturalistic base vocabularies that provides the *intramural excellence* feature that is supposed to underwrite the claim to *extramural authority* (see (1b) above). The theorem highlights in a particularly striking way the inadequacy of the Deductive-Nomological approach to understanding explanation. For it shows how *cheaply* we can buy *derivability from a recursively axiomatizable* (first-order) *theory*, and just how unilluminating that can be. But by doing that, it also highlights the challenge fans of a *naturalistic trans-domain hegemony inference* face in specifying the sort of intramural explanatory excellence or privilege they aspire to demonstrate for some naturalistic base language. How can the relevant notion of explanation, and hence *completeness* (or closedness) of explanation (cf. Lewis's "true and exhaustive account") satisfactorily be made out?

5. **Perverse reverse supervenience** of *any* vocabulary, including φ, on *normative semantic* vocabulary built upon it. This is an argument that I have *not* seen in the literature:

a) At least any *descriptive* vocabulary (alternative: any vocabulary that permits the expression of *knowledge* claims, hence any *cognitively significant* vocabulary) must support assessment of claims according to two sorts of semantic norms or standards: correct_J and correct_T. The social-perspectival difference between these, as an interpretation of the JTB account of knowledge, transposed on pragmatist lines into

an account of what one is *doing* in *attributing* knowledge to someone. It is a linguistic semantic norm that there is a sense of 'correct' in which it is correct to endorse the claim that *p* if and only if *p*. This is the sense of correctness that goes with truth, rather than with justification. Call it 'correct$_T$' In this sense, I am correct$_T$ to believe that the mass of the universe is large enough that it will eventually collapse gravitationally if and only if the mass of the universe is large enough that it will eventually collapse gravitationally, regardless of whether anyone will ever have enough evidence to be correct$_J$, in the sense of *justified* concerning their views on the issue.

b) The normative semantic facts about correctness$_T$ of claims couched in the language of physics supervene on the physical facts expressed in that same language, plus whatever facts (physical or not) it is in virtue of which our words mean what they do. (A sufficiently radical semantic externalism would hold that if the physical facts were at all different, the meanings of physical expressions would be different. Sellars's weaker view is that the meanings would be different only if the physical *laws*, but not the *contingent* physical facts were different.)

c) But notice that if we consider a sufficiently expressively powerful or complete language of physics—in the sense of one that has the expressive resources to state any physical fact whatsoever, no matter how long such a statement might be (it need not be a conjunction, and might even be an infinite collection of sentences)—then the physical facts will supervene on the (semantic) *normative* facts.

d) Does it really make sense to think of such an expressively complete language of physics?

i. It is difficult to say what the physical facts are without at least implicit appeal to such an ideal language in which they could be stated. Phrases such as "the K facts", for values of K such as 'semantic', 'physical', 'moral', and so on are very hard to assign extensions to without appealing to the K-vocabulary in which they are stated.

ii. Nothing that we care about in terms of the consequences of this perverse or reverse supervenience turns on the language in question being *finitary*.

iii. And we can deal with *cardinality* issues in the same way we do when showing that *substitutional* understandings of quantifiers need not diverge in their inferential consequences from *objectual* understandings of quantifiers. That is, we can consider *arbitrary minimal extensions* of the actual language of physics. Such extensions include *names* for *objects* (e.g. distant electrons) that we don't currently have names for. No one extension need have *all* the new names in it for it to be true that facts that must be stated using such names can be expressed in *some* minimal extension of the current language. We can also allow the addition of new *predicates*. Our current belief is that there is only a finite number of subatomic particles. Theorists like Lewis use this to count as physical *objects* all and only the mereological sums of such particles. Is it the case that the language of real number theory allows us to state all the real-number facts? If so, the continuity of the spaces in which those particles are deployed is no bar.

iv. Even if we had to give up on the coherence of the idea of an *expressively complete* language of physics, still the physical facts that *are* statable in whatever

*in*complete language we have will supervene on the semantic normative facts in that language. And this is still a perverse, reverse supervenience.

e) Then the φ-facts globally supervene on the correct$_T$(φ)-facts, in that there could not be two (X-possible) worlds that were just alike in all their correct$_T$(φ)-facts, but differed in their φ-facts. So the physical (for instance) facts supervene on the semantic normative facts. The physical facts supervene on the semantic normative facts, relative to a semantic normative language built on *whatever* physical language we are implicitly invoking when we talk about "the physical facts". There could be no difference in physical facts without a difference in the semantic normative facts.

i. Unless we are radically wrong about what we are capable of meaning, this just follows from the definition of 'correct$_T$', which codifies an essential feature of what descriptive language is—a constitutive norm of describing.

ii. Indeed, the supervenience of the physical facts on the semantic normative facts is significantly more certain than the converse supervenience of the semantic normative facts on the physical facts. For that direction requires the additional claim that all the facts that determine what our words mean—which includes a lot of norms that are *not* of the correct$_T$ sort, for instance *at least* norms of the correct$_J$ sort—themselves supervene on the physical facts.

f) But we would not want, presumably, to draw any grand metaphysical conclusion about the *ontological basicness* of semantic normative facts, from the *fact* that everything else supervenes on these facts.

i. Q: Why not? A: The phenomenon here seems to be like that of the correlation between the length of the flagpole and the length of its shadow. Although, as it were, counterfactuals run in both directions, there is nonetheless an explanatory asymmetry that ought not to be overlooked.

ii. Want an *asymmetric* dependence relation, and for this case, anyway, we have either symmetry or an asymmetry privileging the wrong direction of supervenience.

iii. Compare: the length of the flagpole and the length of its shadow. Mere counterfactuals won't distinguish the dependences, since one can *only* change the shadow-length by changing the pole-length (or changing the whole set-up). What seems to be needed is an appeal to a notion of <u>*explanation*</u>. One can explain how one *knows* the pole-length by appeal to the shadow-length, but not why the pole *has* the length it does. But one *can* explain why the shadow *has* the length it does in terms of the pole having the length it does. And here we can explain *why* there is that explanatory relation, because we can tell a story about an intervening mechanism: how photons moving in straight lines are occluded by the pole and cast the shadow. *If* the photons moved otherwise, or *if* we interfered with them, the shadow *would* have a different length. What could play a corresponding role as symmetry-breaking in the case of the perverse reverse supervenience?

6. **Horgan**'s arguments that global supervenience is too weak to be a form of naturalism worth having. (Q: Does he mean as a particular or a general naturalist claim? A: I don't think it matters, since his arguments would apply to both.):

a) Global supervenience does not rule out supervenient ectoplasm, subject only to its own laws, completely unpredictable on the basis of φ-facts, subject only to the constraint that there is no ectoplasmic difference without a φ-difference. That the

presence of spooks must involve *some* physical difference is not enough to keep them from being spooky. (Q: Is what we are really worried about a failure of explanatory-causal *completeness* of φ-facts?)

b) Global supervenience is compatible with the *only* difference between a world in which some creatures on the Earth have, say, minds, and a world in which none do being a slight displacement of a single physical sub-atomic particle somewhere outside the light-cone of all life on Earth. (Crane elaborates this argument.) But, once again, that would remove the ψ-facts entirely from any *explanatory* connection with φ-facts. It would render them completely *mysterious*. This is an argument originally due to Kim, but Horgan is clearer about it, I think. Stalnaker discusses Kim's response: insist that "very (or sufficiently) similar" φ-worlds be "very (or sufficiently similar)" in their ψ-facts. Horgan concludes that supervenience is not enough. [The scare quotes in this formulation mark an unexplained theoretical parameter, hence an explanatory promissory note.] We need more, we need (in Lycan's happy phrase) super*duper*venience in order to have a naturalism worth having (a naturalism in either the particular, ground-level sense, or in the general, metaphysical sense of (1) above).

7. **Intrinsic vs. Relational** Properties:

a) If that superdupervenience takes the form of *local* or *regional* supervenience, as Horgan is inclined to recommend (note that Kim goes in another, if related, direction: to seriously *disjunctive* reductions in the sense of type-type identities, practically unwieldy but available *in principle*, and supporting not only *definability* but *derivability* of laws, at least *in principle*) then one must be able to distinguish the subset of ψ-properties *of a region r* that supervene on some subset of φ-properties *of that region*. And Horgan himself wants, in addition to local or regional supervenience (on the side of *definability* in classical Carnap-Nagel reducibility), a stronger *explanatory* relation between the vocabularies (on the side of *derivability* in classical Carnap-Nagel reducibility). But I will only consider the first part of his view (in part because little is said here about the other half).

b) In making the claim of regional supervenience, if *all* the ψ- and φ-properties are considered, including the *relational* ones, then since *any* ψ- or φ-difference *anywhere* changes the *relational* ψ- and φ-properties of *every* region (example: Shifting the position of a sub-atomic particle outside the light-cone of life on Earth changes the *physical* properties of everything *within* that light-cone. For being such and such a distance-time removed from a physical event is itself a physical property, albeit a relational one.), *regional* supervenience would collapse into *global* supervenience. But it is intended to be a *stronger* thesis (both in its particular, ground-level, and in its general, metaphysical forms).

c) The conclusion is that in order to have a substantive form of regional supervenience, one must distinguish *intrinsic* from *relational* properties. Then (and only then) one could claim that the *intrinsic* ψ-properties of any suitable region *r* supervene on the *intrinsic* φ-properties of that same region, in the sense that it is not (X-) possible for there to be a difference in intrinsic ψ-properties of any suitable region r ('suitable' since we may want to restrict the claim to, say, whole persons) without a difference in *intrinsic* φ-properties of that same region.

d) It is a philosophical challenge to make out this difference between *intrinsic* and *relational* properties. As we saw with the distinction between *disjunctive* and *non-disjunctive* properties, it cannot be made out solely on syntactic grounds. "…is a parent" (not to be confused with "…is apparent"!) is syntactically a monadic predicate, but it encodes a relational property. And any property P can be expressed as a disjunction of conjunctions: $P \approx P\&Q \vee P\&\sim Q$.

e) Further, this issue is closely related to the issue of how to characterize "Cambridge properties", and how they might be distinguished from "real properties". A typical Cambridge property (so-called by Geach, thinking of McTaggart) would be having the old-Provo eye-color—that is, having the same eye-color as the oldest living inhabitant of Provo, Utah. Again, Fodor defines any particle as being a 'fridgeon' just in case his fridge is on. So when his fridge turns on, it also turns all the particles in the universe temporarily into fridgeons, and gives every macroscopic physical object the new property of being made of fridgeons. For a while there was a small philosophical industry devoted to trying to distinguish 'Cambridge changes' from real ones. I think we have come to see that this enterprise was a misguided one. For any complex relational property such as being a fridgeon or having old-Provo-colored eyes, we can describe *some* inferential circumstances (however outré) in which the credentials of some significant claim would turn precisely on the presence or absence of that property. If we give up the idea of distinguishing 'real', 'significant', or 'natural' relational properties from other relational properties, what are the prospects for distinguishing relational properties from 'intrinsic' ones?

f) I think the state-of-the-art answer to this question (that is, the best proposal currently on the table) is Langton & Lewis (not to be confused with Lewis & Langford, the classic early twentieth century work on modal logic) 1998 *PPR* piece "Defining 'Intrinsic'". It is best approached in stages:

i) Kim (and Peter Vallentyne) proposed calling a property (say, being round) 'intrinsic' iff it could be possessed by something even if that thing were the only thing in the universe, unaccompanied by any contingent object wholly distinct from itself. Call such an object 'lonely' or 'unaccompanied'.

ii) Lewis realized that this definition can't do the work it is intended to do, because being *lonely* turns out to be an intrinsic property on this account, and surely it is in fact a (modal) relational one. He thought nothing like this could work, and tried something else ("Extrinsic Properties", *Phil. Studies* 1983).

iii) Rae Langton realized (in her 1997 Princeton dissertation on Kant) that a definition along these lines could be fixed to get around this sort of counterexample. Just insist that the object be able to have the property unaccompanied *or* accompanied. Slightly more carefully, insist that all four possibilities can arise: the object can have the property either accompanied or unaccompanied, and it can lack the property accompanied or unaccompanied.

iv) Unfortunately, as Lewis and Langton realize, this obviously won't work for *disjunctive* properties. They say:
"Consider the property of being either cubical and lonely, or else non-cubical and accompanied. This property is surely not intrinsic. Yet

having or lacking it is independent of accompaniment or loneliness: all four cases are possible." [p. 335]
And negations of disjunctive properties such as this one also wrongly get counted as intrinsic.

v) Lewis & Langton punt at this point. They just assume that the distinction between *disjunctive* and *natural* properties can be made out *somehow* (One of the ideas they float is that the 'natural' properties are those that play a certain distinctive role in our reasoning, e.g. figuring in *laws*. To make *that* work one would need to solve *first* the problem of distinguishing 'natural' from 'Cambridge' properties), and restrict their definition to those that are not disjunctions of (conjunctions of) natural properties and yet not themselves natural properties.

g) I don't think we need to worry about these further details, because of the philosophical issues that arise already with the possibility-of-loneliness test. What kind of *objects* can coherently be envisaged as existing as the only contingent object in some possible world?

- Kant asks us to imagine two worlds, one consisting only of a left hand, and the other only of a right hand. His point—about the points of view from which these are, and the points of view from which they are not, distinguishable—could be made with a tetrahedron whose vertices are labeled: for instance, with a carbon atom with four different kinds of groups attached to it (which is why organic molecules have enantiomers). But is it really possible for there to be a world consisting only of a *hand*? Think of the reasons that led Aristotle to say that a detached human hand is not a human hand. This hand is, by hypothesis, to be unaccompanied in the whole possible world, past and future included. So there have never been humans in it. It has never been alive. Its component carbon and oxygen atoms were not formed in the interiors of stars early in the history of the universe, since there never were any such stars.
- To change the example slightly, could there be a universe consisting only of a single *cat*? What would make it a *cat*, not having been born of cats, or otherwise having its DNA derivative from genuine cat DNA? Is there in fact any kind of biological organism that we can be confident can coherently be supposed to exist unaccompanied in some possible world?
- Davidson's swamp-man example (cf. sunburn, which is an essentially *relational* phenomenon) suggests that possessing intentional states is an essentially *relational* phenomenon. But now it may be that essentially all *biological* terms and predicates are covertly relational. That would mean that no such things could exist unaccompanied in any possible world.

h) This definition of intrinsicness, in other words, may restrict us to *physical* objects and properties. *Those*, it seems, can exist either accompanied or unaccompanied.

i) Or can they? This is a matter of physics. Perhaps there can be unaccompanied electrons—though physical principles such as the Pauli Exclusion Principle at least relate them to other *possible* inhabitants of the same world. And the curvature of the space-time they inhabit will be different in the accompanied case than it will in the unaccompanied one. But perhaps that does not affect the identity of the lonely electron. For instance, the fact that physics typically makes use (since Newton's

days) of *isolated* systems does *not* mean that those systems being *unaccompanied* is *physically possible*. For many of them, the laws of physics *preclude* their isolation. These are *idealizations* (cf."variable reduction") that help us think *grossly* about physical behavior. The utility of these idealizations in no way turns on their being actually *physically-nomologically possible.* But in any case, let us suppose that a lot of fundamental physical objects and properties *can* coherently be envisaged as unaccompanied.

j) Conclusion: Intrinsicness in the L&L sense drives us down to the level of *fundamental physics* (which might give us only constituents), and even there it is a serious *empirical* question (at any rate, it turns on the details of the physics) whether and when this notion makes sense.

k) But if already at the *biological* level, never mind the *psychological*, *social*, or *normative* levels, a demand for intrinsicness in the sense of the possibility of *unaccompanied* existence rules out the very existence of the objects, never mind their possession of the properties constitutive of this sort of being, then won't we be begging the question in favor of a general metaphysical physicalism (cf. (1) above) if we formulate our question about regional supervenience in terms of a notion of <u>*duplicates*</u> of the sort Lewis & Langton recommend: two things are *duplicates* just in case they have the same intrinsic properties. For, I am claiming, all intrinsic properties *are* physical properties. Higher-level properties, for instance, all *functional* properties, are in this sense *relational* (extrinsic). A physical duplicate of a hand is not a hand. Should we conclude that there 'really' are no hands? There is a danger of the deck being stacked, or the question being begged here. That would be so if *all intrinsic* properties supervened on intrinsic *physical* properties, just because all *intrinsic* properties *are physical* properties.

l) One way to think about the issue is that in restricting our attention to the *intrinsic* properties in the Lewis&Langton sense, we are trading in the *identity* of the objects we are concerned with (e.g. biological ones such as *humans*, never mind psychological ones such as *persons*) for their *physical constitution*. And here the basic lesson to keep in mind is that identity≠constitution. Lumpl (the name for the lump of clay the statue is formed from) ≠ statue. Lumpl≠statue because they have different *modal* properties: Lumpl *would* still exist if the statue were crushed into a sphere, but the statue *would* not. (Q1: Are *modal* properties *intrinsic* properties? It seems not, for what *would happen next* to the molecules in a cat's lungs in an unaccompanied world and in this world is *quite* different. Q2: Is L&L intrinsicness conserved by mereological combination? Constitution is, but mereology allows "disjunctive objects", and so disjunctive properties.) The worry, then, is that in asking whether there can be two objects that are intrinsic φ-duplicates but are not *intrinsic* ψ-duplicates, we have implicitly and covertly restricted ourselves already to the *matter* that *constitutes* the ψ-things, rather than the ψ-things themselves. It would then be no surprise to discover that the *matter constituting* ψ-things supervenes on their intrinsic physical properties. But that is a strategy for buying naturalism too cheaply for what one gets to be worth having.

m) So regional or local supervenience (of either the particular, ground-level or the general, metaphysical kind) too, is a hard thesis to formulate so as to be both *true* and *interesting*.

8. **Stalnaker**:

a) Stalnaker sorts out modally strong/weak supervenience and global/local supervenience.

b) Then he distinguishes two intuitive notions of supervenience, one a kind of liberal reduction, the other a kind of substantive metaphysical dependence, which he thinks stand in a kind of tension.

c) One reason to read this is to point to the Paull and Sider (1992 *PPR*) "In Defense of Global Supervenience" paper, following up on Petrie's (1987 *PPR*) "Global Supervenience and Reduction". Along important dimensions, Stalnaker sums up and supersedes this bit of the literature.

d) "Kim has shown that if A strongly supervenes on B, then every A-property is necessarily equivalent to a property definable in terms of B-properties." [228]. This is done *not* by using the Beth result to turn implicit definitions into explicit ones in the case of *global* (strong) supervenience, but by exploiting the token-token identities that *merely* strong local or regional supervenience is committed to in order to form definitions as possibly infinite disjunctions—which are, accordingly of no *explanatory* use. But if one combines this result of Kim's with the result Stalnaker proves in his Appendix, one gets the same conclusion we can draw from the (much more powerful, but harder to prove) Beth theorem: a sense in which global supervenience entails reducibility in the sense of definability. [See below.]

e) "I think a materialist is committed, in virtue of his materialism, to no supervenience thesis stronger than the thesis that the mental properties globally supervene on physical properties." [228]

f) "Jaegwon Kim has a different reason for thinking that the global supervenience of the mental on the physical may not be sufficiently strong for materialism. The problem is that global supervenience of the mental on the physical is compatible with large and important mental differences being dependent on trivial and seemingly irrelevant physical differences." [229] This is the second of the objections Horgan offers.

g) RS assesses this argument like this: "But sensible materialists are not only materialists, they are also sensible; one should not define materialism so that there cannot be silly versions of it." [229]

h) "Kim's response to his wayward atom example is to suggest a strengthening of global supervenience that requires not only that B-indiscernible worlds be A-indiscernible, but also that worlds that are very similar with respect to the distribution of B-properties be very similar with respect to the distribution of A-properties." [230]

i) RS then turns to the sort of *necessity* involved.

j) Appendix Argument:

Definitions:

(1) "A *strongly supervenes* on B iff for any worlds *w* and *z*, and for any objects *x* and *y*, if *x* has in *w* the same B-properties that *y* has in *z*, then *x* has in *w* the same A-properties that *y* has in *z*."[7]

(3) A *globally supervenes* on B iff any two possible worlds that are B-indiscernible are also A-indiscernible.

Proof:

The claim to be proven is that the global supervenience of A on B is equivalent to the strong supervenience of A on B′, where B′ is the set of properties definable in terms of B properties, in an infinitary language with identity, quantifiers and truth functional operators. It is obvious that strong supervenience implies global supervenience, and that the global supervenience of A on B is the same as the global supervenience of A on B′. What needs to be shown is that if A globally supervenes on B, then A strongly supervenes on B′.

First, define a *complete B-description* of a world *w* as follows: if there are n members of the domain of *w*, the description will begin with n existential quantifiers. If variable *x* corresponds to individual **a**, then the description will contain a conjunct F*x* for each of **a**'s B-properties F, and a conjunct ~F*x* for each B-property that **a** lacks. The description will also include conjuncts $x \neq y$ for each pair of distinct variables *x* and *y* that are bound by the existential quantifiers, and a universal generalization saying that everything is one of the n things. Obviously, any two worlds that have the same complete B-description will be B-indiscernible with respect to a mapping of the domain of one onto the domain of the other.

Now suppose that A globally supervenes on B. Let *w* and *z* be any two possible worlds, and **a** and **b** any two individuals from the domains of *w* and *z*, respectively, such that **a** has all the same B′-properties in *w* that **b** has in *z*. Let ϕ be the complete B-description of *w*, and let *x* be the variable in the description that corresponds to **a**. Drop from ϕ the quantifier that binds *x*, and the result is an open sentence with one free variable that expresses the maximal B′ property that **a** has in *w*. Since **b** has the same B′-properties in *z* as **a** has in *w*, it follows that **b** has this property in *z*. But then the existential generalization of this open sentence, which is equivalent to ϕ, is true in *z*, and so *w* and *z* are B-indiscernible, relative to a mapping that maps **a** to **b**. So since we are assuming that A globally supervenes on B, it follows that *w* and *z* are also A-indiscernible, relative to the same mapping. So **a** has all the same A-properties in *w* that **b** has in *z*. Therefore, A strongly supervenes on B′.

Notes on this argument:

k) This is a strengthening of the Kim result mentioned in (4) above.

l) It works basically by creating a B' property: being such that everything else in the world has exactly the B-properties that it has. This encodes into a property of *a* all the B properties of its world.

m) Notice that this is why we need an *infinitary* language, which makes this property of no use at all in any kind of explanation.

n) Notice further that this is wholly a *model-theoretic* result, *not* one that starts with *possible worlds*. For possible worlds don't come with *domains*, i.e. the set of all the *objects* in the world, which can be enumerated so we know there are exactly *n* of them. It is *algebraic relational structures* that have *domains*. This notion does not even make sense for possible worlds in general. For one cannot count *objects*. 'Object' is a pseudo-sortal, which does not come with a principle of individuation. One can count people, or electrons, but 'object' in effect quantifies over, or stands in

for, a whole *set* of sortals—allowing *any sortal at all*. But this is *not* a well-defined set for possible worlds in general. How many *objects* are there on my desk? Do the shadows count? Is the upper-left curved portion of the shadow an *object*? Is every spatial part of my pen an *object*? David Lewis thinks he can count objects, because he *stipulates* that the *basic* objects are sub-atomic particles [recall my rant about philosophers' toy physics], and that *all* the objects in any world are all and only the mereological sums of those basic objects. This uses one (or perhaps a finite set) of particle-sortals, and mereology. But going *that* way is question-begging in the context of a general discussion of supervenience. So this appeal to model-theory might be laid alongside the Beth theorem, and the considerations Etchemendy raises about the collision of the model-theoretic and possible worlds frameworks in thinking about the concept of logical consequence.