3

Artificial Intelligence and Analytic Pragmatism

1 AI-functionalism

The thought with which I introduced meaning-use analysis, and the paradigm of a pragmatically mediated semantic relation, arises when we put together two sorts of story:

- an account of what one must *do* in order to count as *saying* something—that is, of some practices-or-abilities that are PV-sufficient to deploy a vocabulary, and
- a characterization of another vocabulary that one can use to *say* what it is one must *do* to be doing something, for instance, in order to be *saying* something else—that is, of a vocabulary that is VP-sufficient to *specify* the practices-or-abilities, which might be PV-sufficient to deploy another vocabulary.

When we compose these, the resultant meaning-use relation (MUR) is the relationship between vocabularies that I have called the ‘‘pragmatic metavocabulary’’ relation. I have suggested that this relation is most illuminating when the pragmatic metavocabulary is demonstrably *expressively weaker* than the vocabulary for which it is a pragmatic metavocabulary. This is what I have called ‘‘pragmatic expressive bootstrapping,’’ in the strict sense. We have seen several examples of this phenomenon:

- *Syntactic* pragmatic bootstrapping, within the Chomsky hierarchy of grammars and automata, in that expressively weaker *context-free* vocabularies are VP-sufficient to specify Turing machines (two-stack push-down automata), which are in turn PV-sufficient to deploy (produce and recognize) expressively stronger *recursively enumerable* vocabularies.

- I argued that *non*-indexical vocabulary is VP-sufficient to specify practices PV-sufficient to deploy *indexical* vocabulary.
- I have mentioned, though not discussed, Huw Price's pragmatic naturalism, which denies the semantic reducibility of normative to naturalistic vocabulary—and even the supervenience of the one on the other—but which seeks to lessen the sting of that denial by specifying in a naturalistic vocabulary what one must *do* in order to deploy various irreducibly non-naturalistic vocabularies, for example normative or intentional ones.

I will argue in later lectures that deontic *normative* vocabulary is a sufficient pragmatic metavocabulary for alethic *modal* vocabulary: a case where the expressive ranges are at least impressively *different*, even if not rankable as strictly expressively weaker and stronger.

In this lecture, I will discuss another philosophically significant contention of this kind: the claim, thesis, or program that is usually associated with the rubric "artificial intelligence." Very crudely, AI is the claim that a computer could in principle *do* what is needed to deploy an autonomous vocabulary, that is, in this strong sense, to *say* something. It is accordingly a thesis about meaning-use relations, in my sense. The classical Turing test for the sort of 'intelligence' at issue is a *talking* test; something passes it if, by talking to it, one cannot tell it from a human speaker, that is, from someone who engages in autonomous discursive practices, someone who deploys an autonomous vocabulary. 'Intelligence' in this sense just consists in deploying such a vocabulary. Classical AI-functionalism is the claim that there is some computer program (some algorithm) such that anything that runs that program (executes that algorithm) can pass the Turing test, that is, can deploy a vocabulary in the sense in which any other language-users do. And that is to say that a computer language, in which any such algorithm can be expressed, is in principle VP-sufficient to specify abilities that are PV-sufficient to deploy an autonomous vocabulary. **So in my terms, classical AI-functionalism claims that computer languages are in principle sufficient *pragmatic metavocabularies* for some autonomous vocabulary**. (Did you see that coming?)

Now I take it that computer languages are not themselves autonomous vocabularies. For such context-free languages lack essential kinds of vocabulary. We cannot make sense of linguistic communities that speak only Prolog or C++ (though some groups of engineers, when talking among themselves, on occasion seem to come close). Insofar as that is right, the basic claim of AI-functionalism is an *expressive bootstrapping* claim about computer languages as pragmatic metavocabularies for much more expressively powerful vocabularies, namely natural languages. Of course, AI has not traditionally been thought of as an expressive bootstrapping claim about a pragmatic metavocabulary. How could it have been? But it deserves a prominent place on the list of philosophically significant pragmatic expressive bootstrapping claims I just offered. And it should be a principal topic of philosophical meaning-use analysis. So let us see what the meaning-use analysis metavocabulary I have been deploying can help us understand about it—what lessons these metaconceptual tools can teach us when they are applied to this issue of independent interest.

Although its twentieth-century version developed later than the others, functionalism in the philosophy of mind, including its central computational species, deserves to be thought of as a *third* core program of the classical project of philosophical analysis, alongside empiricism and naturalism. (For reasons indicated in the previous lecture, I think of behaviorism as a larval stage of functionalism.) And since AI-functionalism concerns the relation between practices or abilities and the deployment of vocabularies, insofar as functionalist successors to behaviorist programs in the philosophy of mind do deserve a prominent place at the analytic table, that fact indicates that the sort of broadening of the analytic semantic project to include pragmatics that I have been recommending has in fact implicitly been under way for some time.

2 Classic symbolic artificial intelligence

I take the working-out of various forms of functionalism in the philosophy of mind to have been one of the cardinal achievements of Anglophone philosophy in the second half of the twentieth century. One of the things I think we have found out along the way is that functionalism is a more promising explanatory strategy when addressed to *sapience* than when addressed to *sentience*—when it is addressed to our understanding of states such as belief, rather than pains or sensations of red. In broadest terms, the basic idea of functionalism is to assimilate bits of intentional vocabulary such as ‘‘belief that *p*’’ to terms classifying something in terms of the role it plays in a more complex system. So the relations between ‘belief’, ‘desire’, ‘intention’, and ‘action’ might be modeled on the relations between ‘valve’, ‘fluid’, ‘pump’, and ‘filter’. The most immediate attraction of such an approach is the *via media* it provides between the traditional alternatives of materialism and dualism. All valves, that is, all things playing the functional role of a valve in any system, are physical objects, and they can function as valves only in virtue of their physical properties. So far, *materialism* was right: functional vocabulary applies exclusively to physical objects. But what valves have in common that makes that term properly apply to them is not a *physical* property. Mechanical hydraulic valves, heart valves, and electronic valves may have no physical properties in common that they do not share with a host of non-valves. So far, *dualism* was right: functional properties are not physical properties. *Automaton* functionalism is a species of this general view that looks specifically at the functional roles items can play in multi-state transducing automata. By the term ‘AI-functionalism’ I shall mean automaton functionalism about *sapience*—about what it is in virtue of which *intentional-state* vocabulary such as ‘‘believes that’’ is applicable to something, that is, in the terms I have been using (and which are endorsed by appeals to the Turing test), the capacity to engage in any autonomous discursive practice, to deploy any autonomous vocabulary, to engage in any discursive practice one could engage in though one engaged in no other.

So understood, AI-functionalism admits of different interpretations. Approaching it as asserting a particular kind of pragmatically mediated semantic relation between vocabularies—as making an expressive bootstrapping claim about a particular kind of pragmatic metavocabulary for some autonomous vocabulary—as meaning-use analysis suggests, leads to a characterization that is in important ways broader than traditional formulations. I want to begin by saying something about that difference.

I will call what I take to be the received understanding of the central claims of AI, what John Searle calls the ‘‘strong thesis of AI,’’ ‘‘*classical symbolic AI*’’—or sometimes ‘‘*classy* AI,’’ for short. Here is how I understand it. Its slogan is: ‘‘*Mind is to brain as software is to hardware.*’’ It sees a crucial difference between modeling the mind on computer programs and all previous fashionable, rashly enthusiastic claims that some bit of impressively powerful new technology would also, *inter alia*, give us the key to unlock the secrets of the mind—telephone switchboards, clockworks, and, if we go far enough back, even potters' wheels having been taken to play that role. For computing is manipulating symbols according to definite rules (the algorithms implicit in automaton state-tables). And, the claim is, thinking or reasoning, the fundamental sort of operation or activity that constitutes sapience, just *is* manipulating symbols according to definite rules. This *computational theory of the mind* is the basis of the standard argument for AI-functionalism. It is a view that long antedates the advent of computers, having been epitomized already by Hobbes in his claim that "reasoning is but reckoning."

Now the plausibility of understanding thinking as symbol-manipulation at all depends on taking symbols to be more than just sign-designs with a syntax. They must be *meaningful, semantically* contentful signs, whose proper manipulation—what it is *correct* to do with them—depends on the meanings they express, or on what they represent. Traditionally, this fact meant that there was a problem reconciling the computational view of the mind with naturalism. Physics does not find meanings or semantic properties in its catalogue of the furniture of the world. They are not, or at any rate not evidently, physical properties. So how could any physical system *be* a computer—a *symbol*-manipulator in the relevant sense—and so respond differentially to signs depending on the *meanings* they express? Looking back from the vantage point vouchsafed us by the development of actual computing machinery—and the realization that doing numerical calculation by the algorithmic manipulation of numerals was only one instance of a more general symbol-manipulating capacity—provides a possible answer. Already for Descartes, the thoroughgoing isomorphism he had established between algebraic formulae and geometric figures suggested that manipulating the formulae according to the rules proper to them could not just *express*, but also *constitute* or *embody*, an understanding of the figures. The isomorphism amounts to an encoding of *semantic* properties in *syntactic* ones. A physical system can accordingly be a computer—manipulate symbols in ways that accord with their meanings—because such an encoding ensures that, in Haugeland's slogan, if the automaton takes care of the *syntax*, the *semantics* will take care of itself.[1]

Usually, though, what you get when you manipulate symbols in ways that exploit isomorphisms to what they are symbols of is a *simulation*. Computers

[1] This characterization of classical symbolic AI owes a lot to John Haugeland's *Artificial Iintelligence: The Very Idea* (MIT Press, 1989). My own thoughts on this subject were worked out in the course of teaching undergraduate AI courses based on this text.

can manipulate symbols to model traffic patterns, weather systems, and forest fires. No one is liable to confuse the symbol-manipulating with the phenomena it simulates—the computation with the traffic, the weather, or the fire. But AI-functionalism claims that, unlike these cases, manipulating symbols in ways that suitably respect, reflect, and exploit isomorphisms with what those symbols for that very reason count as expressing or representing is not just a *simulation* of thinking, but *is* thinking itself. That is what it *is* to deploy a vocabulary *as* a vocabulary, that is, as meaningful. The *only* reason for according thought this uniquely privileged position—as the one phenomenon that cannot be symbolically *simulated* without thereby being actually *instantiated*—is whatever reason there is to think that the symbolic-computational theory of the mind is correct. And that is a very substantive, potentially controversial theory of sapience, with a correspondingly large burden of proof.

3 A pragmatic conception of artificial intelligence

I think that symbolic AI's focus on the Turing test is appropriate. There just is no point in insisting that something that is genuinely indistinguishable (including, crucially, dispositionally counterfactually) from other discursive practitioners in conversation—no matter how extended and wide-ranging in topic—should nonetheless not be counted as *really* talking, so thinking (out loud), and deploying a meaningful vocabulary. But although the slide can seem unavoidable, it is a long way from acknowledging the criterial character of the Turing test for sapience to endorsing the computational theory of the mind on which classical symbolic AI is predicated. The line of thought I have just rehearsed invites a focus on the issue of the *symbolic* character of thought that I think is ultimately misleading. And for that reason it *mis*locates, as it seems to me, what really is the most important issue in the vicinity: the claimed *algorithmic* character (or characterizability) of thought or discursive practice.

In Lecture 2 I argued that, from the point of view of meaning-use analysis, the principal significance of automata does not lie in their capacity to manipulate symbols, but rather in their implementing a distinctive kind of PP-sufficiency relation. Multi-state transducing automata *algorithmically elaborate* a set of primitive abilities into further abilities—abilities which, just because they can be so exhibited, can then be regarded as complex, as pragmatically *analyzable* into those primitive abilities plus the basic algorithmic elaborating abilities. This characterization of automata suggests that AI be understood broadly as a claim to the effect that such an analysis or decomposition is possible of some autonomous discursive practice—the practice-or-ability to deploy some vocabulary that can be deployed though one deploys no other. That is, it claims that some autonomous discursive practice can be exhibited as the algorithmic elaboration of a set of primitive abilities, which are accordingly PP-sufficient for that autonomous discursive practice.

That claim by itself would not be interesting or controversial. For the null elaboration is also an algorithmic elaboration (albeit a degenerate one). So the condition would be trivially satisfied, just because there *are* autonomous discursive practices-or-abilities. What is needed to turn the claim that some set of primitive abilities can be algorithmically elaborated so as to be PP-sufficient for some autonomous discursive practice into a genuinely substantive claim is a further constraint on the primitive abilities. Given the reasons for being interested in AI-functionalism in the first place, what we want is to stipulate that what are to be counted as primitive abilities with respect to such an algorithmic elaboration must not themselves in some sense already be *discursive* abilities.

Here is the version that I propose. What I will call the "algorithmic pragmatic elaboration" version of AI-functionalism—or just "pragmatic AI"—is the claim that there is a set of practices-or-abilities meeting two conditions:

1. It can be algorithmically elaborated into (the ability to engage in) an autonomous discursive practice (ADP).
2. Every element in that set of primitive practices-or-abilities can intelligibly be understood to be engaged in, possessed, exercised, or exhibited by something that does *not* engage in any ADP.

In the terminology of meaning-use analysis, the first of these is a kind of PP-sufficiency claim—specifically, an algorithmic elaboration PP-sufficiency claim. The second is the denial of a set of PP-necessity claims.

This approach to AI-functionalism shifts the focus of attention away from the role of *symbols* in thought, away from the question of whether thinking just *is* manipulation of symbols, and away from the issue of whether isomorphism is sufficient to establish genuine ('original', rather than merely 'derivative') semantic contentfulness. It is true that I am here still thinking of what is at issue in sapience as a matter of deploying *vocabularies*, that is, using symbols, semantically significant signs—not in a derivative way, but in whatever way is fundamental in the sense of being exhibited by *autonomous* discursive practices-or-abilities, and the vocabularies they deploy. But—and here is the important difference from classical symbolic AI—the connection to computers (or as I would prefer to say, *automata*) is established not via the principle that computers are symbol-manipulating engines and that, according to the computational theory of the mind, thinking just consists in manipulating symbols, but rather via PP-sufficiency of the algorithmic elaboration sort that I discussed in Lecture 2. And the structural question AI-functionalism asks is an issue that can arise for *any* ability—not just those that involve symbol use. That is, for *any* practice-or-ability *P*, we can ask whether that practice-or-ability can be algorithmically *decomposed* (pragmatically analyzed) into a set of primitive practices-or-abilties such that:

1. they are PP-sufficient for *P*, in the sense that *P* can be algorithmically elaborated from them (that is, that *all* you need in principle to be able to engage in or exercise *P* is to be able to engage in those abilities plus the algorithmic elaborative abilities, when these are all integrated as specified by some algorithm); and
2. one could have the capacity to engage in or exercise *each* of those primitive practices-or-abilities without having the capacity to engage in or exercise the target practice-or-ability *P*.

If those two conditions are met, we may say that *P* is ***substantively*** *algorithmically decomposable* into those primitive practices-or-abilities. So, for instance, the capacity to do long division *is* substantively algorithmically decomposable, into the primitive (with respect to this decomposition) capacities to do multiplication and subtraction. For one can learn to multiply, or again, to subtract, without yet having learned how to divide. Perhaps (though I doubt it) the capacity to play the piano is like this, since one can learn how to finger each key individually, and to adjust the intervals between doing so. By contrast, the capacities to respond differentially to red things and to wiggle my index finger probably are not substantively algorithmically decomposable into more basic capacities. These are not things that I do *by* doing something else. If I do not have those abilities, there is no way to put them together as the complex results of some structured sequence of other things—even with the flexibility of conditional branched schedule algorithms, hence of Test-Operate-Test-Exit feedback loops of perception, action, and further perception of the results of the action. The abilities to ride a bicycle, to swim, or to hang-glide might or might not be substantively practically algorithmically decomposable, and the empirical question of whether they are, and if so, how, is of considerable pedagogical significance (about which more later).

So the question of whether some practice-or-ability admits of a substantive practical algorithmic decomposition is a matter of what contingent, parochial, matter-of-factual PP-sufficiencies and necessities actually are exhibited by the creatures producing the performances in question. That question is very general and abstract, but also both empirical and important. It is a very general *structural* question about the ability in question. That issue as such, however, has *nothing whatever* to do with *symbol* manipulation. My suggestion is that we think of the core issue of AI-functionalism as being of this form. The issue is whether *whatever* capacities constitute sapience, *whatever* practices-or-abilities it involves, admit of such a substantive practical algorithmic decomposition. If we think of sapience as consisting in the capacity to deploy a vocabulary, so as being what the Turing test is a test for, then since we are thinking of sapience as a kind of symbol use, the target practices-or-abilities will also involve symbols. But that is an entirely separate, in principle independent, commitment. That is why I say that classical symbolic AI-functionalism is merely one species of the broader genus of algorithmic practical elaboration AI-functionalism, and that the central issues are mislocated if we focus on the *symbolic* nature of thought rather than the substantive practical algorithmic analyzability of whatever practices-or-abilities are sufficient for sapience.

4 Arguments against AI-functionalism: ranges of counterfactual robustness for complex relational predicates

Because the two stand or fall together, arguments against the plausibility of the claims of classic symbolic AI-functionalism usually take the form of arguments against the computational theory of the sapient mind.

These arguments include doubts about the possibility of explicitly codifying in programmable, hence explicitly statable, *rules* all the implicit practical background skills necessary for thoughtful engagement with the world, challenges to the adequacy of the semantic epiphenomenalism inherent in treating syntactic isomorphism as sufficient for non-derivative contentfulness, and reminders of the sort epitomized by Searle's Chinese Room thought-experiment[2] of how badly the essentially third-person point of view of this sort of functionalist successor to behaviorism fits with intuitions derived from our first-person experience of understanding, grasping meanings, deploying vocabularies, and having contentful thoughts. Reasons for skepticism about the sort of AI understood instead as claiming the substantive algorithmic decomposability of autonomous discursive practices-or-abilities into non-discursive ones must take a distinctly different shape.

For instance, Dreyfus objects to classical symbolic AI on the grounds that it requires that all the implicit practical skills necessary for understanding our ordinary life-world have to be made explicit in the form of rules (codified in programs).[3] He diagnoses classy AI as built around the traditional platonist or intellectualist commitment to finding some bit of explicit knowing- (or believing-)*that* behind every bit of implicit practical knowing-*how*. Like Dewey, he is skeptical about that framing commitment. By contrast, the corresponding argument against the substantive practical algorithmic decomposability version of AI would have to offer reasons for pessimism about the possibility of algorithmically resolving essentially discursive knowing- (or believing-)*that* without remainder into non-discursive forms of knowing-*how*. **Whatever problems there may be with this kind of AI, they do not stem from some hidden *intellectualism*, but, on the contrary, concern the particular variety of *pragmatism* it articulates: algorithmic pragmatism about the discursive**. For what makes the substantive algorithmic practical elaboration model of AI interesting is the relatively precise shape that it gives to the *pragmatist* program of explaining knowing-that in terms of knowing-how: specifying in a non-intentional, non-semantic vocabulary what it is one must *do* in order to count as deploying some vocabulary to say something,

[2] In "Minds, Brains, and Programs" (1980), reprinted in John Haugeland (ed.), *Mind Design II* (MIT Press, 1997).

[3] For instance, in Hubert L. Dreyfus, *What Computers Still Can't Do* (MIT Press, 1997).

hence as making intentional and semantic vocabulary applicable to the performances one produces (a kind of pragmatic expressive bootstrapping).

What arguments are there against this pragmatist version of AI? The form of the claim tells us that to argue against the practical algorithmic elaboration version of AI we must find some aspect exhibited by all autonomous discursive practices that is not algorithmically decomposable into non-discursive practices-or-abilities. That would be something that is PV-necessary for deploying any autonomous vocabulary (or equivalently, PP-necessary for any ADP) that cannot be algorithmically decomposed into practices for which no ADP is PP-necessary.

I do not claim to have a knock-down argument here. But the best candidate I can think of to play that role is the practice of doxastic updating—of adjusting one's other beliefs in response to a change of belief, paradigmatically the addition of a new belief.

It is pretty clear that this set of practices-or-abilities is a PV-necessary aspect of the deployment of any vocabulary. For any set of practices to count as *discursive*, I claimed last time, it must accord some performances the significance of *claimings*. It is a necessary feature of that significance that what is expressed by those performances stands to other such contents in broadly *inferential* relations of being a reason for or against. That is, the practical significance of claiming includes undertaking a commitment that has other commitments and entitlements (or lack of entitlements) to commitments as its consequences, that can itself be a consequence of other commitments, and whose entitlement also depends on its relation to one's other commitments. One understands or grasps the content expressed by some bit of vocabulary that can be used to make claims only to the extent to which one can tell in practice (respond differentially according to) what follows from it and what it follows from, what other commitments and entitlements the various commitments it can be used to undertake include and preclude. And that is to say that one understands what a bit of vocabulary means only insofar as one knows what difference undertaking a commitment by its use would make to what else the one using it is committed or entitled to—that is, insofar as one knows how to update a set of commitments and entitlements in the light of adding one that would be expressed using that vocabulary (keeping deontic score). Discursive understanding of this sort is a more-or-less affair. One need not be omniscient about the significance of a bit of vocabulary in order to deploy it meaningfully. But if one has *no* idea what practical consequences for other commitments a claim using it would have, then one associates no meaning with it at all.

If all that is right, then the question of whether doxastic updating can serve as a reason to be pessimistic about the practical algorithmic elaboration version of AI comes down to an assessment of the prospects for a substantive algorithmic decomposition of the ability to update. Why might one think that no such decomposition is possible—that is, that that essential discursive ability could not be algorithmically elaborated from any set of non-discursive abilities? The key point, I think, is that the updating process is highly sensitive to collateral commitments or beliefs. The significance of undertaking a new commitment (or relinquishing an old one) depends not just on the *content* of *that* commitment, but also on what *else* one is already committed to. I will argue in my next lecture that we can think of this global updating ability as a collection of sub-abilities: as the capacity, in one's actual doxastic context, to associate with each commitment a range of counterfactual robustness. To do that is to distinguish, for each commitment (including inferential commitments), which further commitments *would*, and which would *not*, infirm or defeat it. This includes not only claims that are incompatible with it, but also claims that are incompatible with it in the context of one's other collateral beliefs—that is, which complete a *set* of claims that are jointly (but perhaps not severally) incompatible with it.

I take it that there is nothing unintelligible about having such practical abilities, fallible and incomplete though they may be, to distinguish claims that *are* from those that are *not* contextually incompatible with a given claim. And it is clear that a global updating capacity can be algorithmically elaborated from such abilities to discriminate ranges of counterfactual robustness. But I do not think that this sort of ability is a good candidate for an algorithmic decomposition that is *substantive* in the sense I have given to that term. For I do not see that we can make sense of abilities to discern ranges of counterfactual robustness being exhibited, whether severally or collectively, by *non-discursive* creatures. The problem is that the productivity of language guarantees that anything that can talk can form predicates specifying an indefinitely large class of relational properties. As a consequence, any new information about any object carries with it new information of a sort about every other object. For *any* change in *any* property of one changes *some* of the relational properties of *all* the rest. The problem in a nutshell is that doxastic updating for language-users requires distinguishing among all of these, those that *are* from those that are *not* relevant to the claims and inferences one endorses—that is, those which fall within the range of counterfactual robustness of those claims and inferences. And it is *not* plausible, I claim, that *this* ability can be algorithmically decomposed into abilities exhibitable by non-linguistic creatures.

Why not? The logical and computational versions of what the AI community calls the "frame problem" showed that updating requires exercising what turns out to be a crucially important but easily overlooked cognitive skill: the capacity to *ignore* some factors one is capable of attending to. But worrying about the practical engineering problem of how to implement such an ability in finite-state automata revealed a deeper theoretical conceptual problem, which concerns not *how* to ignore some considerations, but *what* to ignore. A simple version of the issue is afforded by the familiar observation that anything is similar to anything else in an infinite number of ways, and also dissimilar to it in an infinite number of ways. For instance, my left little finger and Bach's second Brandenburg concerto are not only different in countless ways, but are similar in that neither is a window-shade, nor a prime number, neither existed before 1600, and both can be damaged by the careless use of stringed instruments. Dealing with objects as knowers and agents requires the ability to *privilege* some of these respects of similarity and difference—to sort the myriad of such respects into those that *are* and those that are *not* relevant to or significant for the inferences, theoretical and practical, to and from the claims about those objects with which one is concerned. In the sort of case I want to focus on, there are lots of complex relational properties that we should usually ignore in our reasoning.

For instance, Fodor defines any particle as being a 'fridgeon' just in case his fridge is on.[4] 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. Again, a death in a distant place can give me the new property of having the same eye-color as

[4] In "Modules, Frames, Fridgeons, Sleeping Dogs and the Music of the Spheres," in Z. Pylyshyn (ed.), *The Robot's Dilemma: The Frame Problem in Artificial Intelligence* (Ablex, 1987).

the oldest living inhabitant of Provo, Utah. Usually I ought to ignore these properties and facts. One of the lessons of the narrower, engineering versions of the frame problem is that updating becomes computationally infeasible if I cannot do that, and am accordingly obliged to check every one of my beliefs and the inferences that support them to see whether they are infirmed by those facts—to be sure that my conclusion that the solid floor will bear my weight is not affected by its suddenly consisting of fridgeons and that my inferential expectation that I will see better if I put on my glasses is still a good one even though my eyes have the new Provo property. For a while there was a small philosophical industry devoted to trying to distinguish what Geach (thinking of McTaggart) called 'Cambridge changes' from real ones.[5] I think we have come to see that this enterprise is 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. What we need to be able to do is not to classify some properties as, in effect, irrelevant *tout court* (irrelevant to what?), but for each inference, to distinguish the considerations that are irrelevant to its goodness, which should accordingly be ignored. This ability is necessary to deal with what Fodor calls epistemological 'isotropy': the fact that any belief is potentially evidentially relevant to any other, given a suitable context of collateral beliefs.

I am claiming that:

- One cannot talk unless one can *ignore* a vast variety of considerations one is capable of attending to, in particular those that involve complex relational properties, that lie within the range of counterfactual robustness of various inferences.
- Only something that can *talk* can do that, since one cannot *ignore* what one cannot *attend* to (a PP-necessity claim), and for many complex relational properties, only those with access to the combinatorial productive resources of a *language* can pick them out and respond differentially to them. No non-linguistic creature can be concerned with fridgeons or old-Provo eye colors.
- So language use, deploying autonomous vocabularies, brings with it the need for a new kind of capacity: for each inference one entertains, to

[5] *God and the Soul* (Routledge, 1969), 71.

distinguish in practice among all the new complex relational properties that one comes to be able to consider, those that are, from those that are not relevant to assessing it.

- Since non-linguistic creatures have no semantic, cognitive, or practical access at all to most of the complex relational properties they would have to distinguish to assess the goodness of many material inferences, there is no reason at all to expect that that sophisticated ability to distinguish ranges of counterfactual robustness involving them could be algorithmically elaborated from the sorts of abilities those creatures do have.[6]

5 Practical elaboration by training

One might reasonably wonder whether, if the sort of argument I have sketched against the substantive algorithmic decomposability of autonomous discursive practices were successful, it would not prove too much. Non-linguistic creatures do, after all, acquire the ability to engage in discursive practices. They do cross the boundary I have been worrying about, and begin deploying vocabularies. This is true both of human infants and, at some point in the past, of our hominid ancestors. The ontogenetic and phylogenetic acquisition of discursive capacities did not, and does not, happen by magic. If discursive practices-or-abilities really are *not* substantively algorithmically decomposable without remainder into non-discursive ones, how *are* we to understand the development of discursive out of non-discursive practices?

I think the answer is that besides algorithmic elaboration there is another, more basic sort of PP-sufficiency relation—another way in which one set

[6] This last claim is a somewhat delicate one. I am *not* using as a premise the claim that we cannot make sense of the possibility of substantively algorithmically decomposing the capacity to be aware of a full range of complex relational properties, by deploying a suitable vocabulary. That is part of the conclusion I am arguing for. I *am* claiming, first, that the ability to ignore the vast majority of complex relational properties that are irrelevant to a given inference in the sense that they fall within its range of counterfactual robustness cannot be taken as *primitive* with respect to a *substantive* algorithmic practical decomposition of discursive practices-or-abilities, and, second, that we have no idea at all how even primitive non-discursive abilities that *could* be substantively algorithmically elaborated into the capacity to form the complex predicates in question could be further elaborated so as to permit the sorting of them into those that do and those that do not belong in the range of counterfactual robustness of a particular inference.

of practices-or-abilities can practically suffice for the acquisition of another. Sometimes those who can engage in one set of practices can learn or be trained to engage in another—not because the target practices can be *algorithmically* elaborated from the original ones, or from some further set into which they can be decomposed, but just because, as a matter of contingent empirical fact concerning creatures of that particular kind, anyone who has the one set of capacities can be brought to have the other as well. So it might be that those who can draw realistic portraits of horses can be brought also to draw realistic portraits of humans, forge signatures, fold origami gracefully, and arrange flowers. If so, no doubt our account of why these other abilities were especially accessible to those who possess the original one would invoke something like "eye-hand co-ordination" or "fine-muscle control." But that is not at all to say that there must be some set of specifiable basic abilities out of which, say, the capacity to draw a good likeness of a friend could be *algorithmically* elaborated. That capacity might admit of no algorithmic decomposition. Certainly the fact that people who can do some other sorts of things can learn or be taught also to do this does not entail or require that there be such a decomposition.

When as a matter of fact there is a course of practical experience or training that will bring those who have one set of abilities to have another set of abilities, I will say that the second can be "practically elaborated by training" from the first. Like algorithmic elaboration, practical elaboration by training is a kind of PP-sufficiency relation. The hallmark of the difference between them is that we can say exactly and in advance what the practices that *implement* PP-sufficiency by algorithmic elaboration are—what *else* besides exercising the primitive abilities one must be able to do in order to elaborate them algorithmically into the target ability. These *elaborative abilities* are things like response substitution and arbitrary state formation—and in general the abilities that suffice to execute a conditional branched-schedule algorithm. These algorithmic elaborative abilities are all that is needed, for instance, to turn the capacity to multiply and subtract into the capacity to do decimal division. And we know how to build machines that have *these* elaborative abilities. By contrast, in the case of practical elaboration by training, we have no idea how to specify in advance the abilities that implement the sufficiency of rote repetition for memorizing the alphabet, or practice for catching a ball or drawing a recognizable face. And where we *can* say something about the abilities that implement PP-sufficiency relations of the practical-elaboration-by-training sort, we find that they both vary wildly from case to case, and depend heavily on parochial biological, sociological, historical, psychological, and biographical contingencies. Finally, where the question of whether one set of well-defined practices-or-abilities can be elaborated *algorithmically* into another is one that can in principle be settled a priori, from one's armchair, the question of whether it is *practically* PP-sufficient for some particular creature or kind of creature, in a particular context, by some specified training regimen, is one that can only be settled empirically.

I think an appreciation of the centrality of this sort of PP-sufficiency relation—which obtains when, as a matter of fact, creatures of a certain sort who can engage in a practice (exhibit an ability) can be brought or can learn to engage in (or exhibit) another—is one of the master ideas animating the thought of the later Wittgenstein. Again and again he emphasizes the extent to which our discursive practices are made possible by the fact that, as a matter of contingent fact, those who have one set of abilities or can engage in one set of practices can be brought by training to exhibit or engage in another. We can be trained to count, associate sounds with written shapes, and respond to signposts, and to exercise those abilities in new cases by "going on in the same way" as others who share our training (and wiring) would. Wittgenstein is, of course, concerned to show us to what extent and in how many ways our discursive practices-or-abilities depend on things that we could not be *taught* to do (by being told) if we could not be *trained* to do them (by being shown). But I think he also sees practical elaboration by training as the principal motor of our discursive practices-or-abilities, as what gives them their *theoretically motley* but *practically tractable* shapes. As I read him, Wittgenstein thinks that *the* most fundamental discursive phenomenon is this way in which the abilities required to deploy one vocabulary can contingently be practically *extended*, elaborated, or developed so as to constitute the ability to deploy some further vocabulary. We may think in this connection of the examples I mentioned in Lecture 1, of the sort of thought-experiments he invites us to conduct concerning this sort of process of *pragmatic projection* of one practice into another: the fact that people who could already use proper names for people could catch on to the practice of using them also for rivers, and that people who could already talk about having gold in their teeth could catch on to talking about having pains in their teeth. The way in which prior abilities are recruited by training in the service of developing new ones is in general unsystematic, not codifiable in rules or algorithms, and not predictable or explicable from first principles. Wittgenstein sees this sort of non-algorithmic practical elaboration as ubiquitous and pervasive. It results in a permanent process of practical discursive mutation that is on the one hand mediated by the productivity of language, and on the other limits its diachronic systematicity.

So the answer to the question with which I began this section is that we do not need to assume that discursive practice is substantively *algorithmically* decomposable into non-discursive practices-or-abilities, on pain of making entering into those practices and acquiring those abilities—by us as a species, and as individuals—unintelligible, because there is another sort of PP-sufficiency relation besides algorithmic elaboration: practical elaboration by training. We need to acknowledge this sort of PP-sufficiency in any case, in order to account for the provenance of the abilities treated as primitive for the purposes of algorithmic elaboration. And Wittgenstein urges us to see this sort of elaboration not only as crucial for the advent of discursive practices-or-abilities, but also as pervasive within up-and-running discursive practices, alongside algorithmic elaboration.

I said at the outset of my story that one of the aims of the sort of analytical pragmatism for which I am seeking to sketch a theoretical basis is to show how Wittgenstein's pragmatist insights need not be taken to underwrite a theoretical quietism antithetical to the project of traditional philosophical analysis, but how those insights can instead be taken on board and pressed into the service of a further pragmatic development and elaboration of that project. Acknowledging the pervasiveness and centrality of non-algorithmic practical elaboration by training need not be the death of theoretical analysis of discursive practice and its relation to the semantic contents expressed by the vocabularies deployed in that practice. For the analytical Wittgensteinian pragmatist, appeal to algorithmically non-decomposable, contingent, parochial abilities is compatible with investigating PP-sufficiency and PP-necessity *dependency* relations between such abilities and practices, as well as the PV- and VP-sufficiency relations they stand in to vocabularies. I would like to close this lecture by outlining one analytic issue that I think is raised directly by the consideration of what I will call *pedagogical* practical elaboration and decomposition of practices and abilities.

I have pointed out that one set of practices-or-abilities can be elaborated into another by a process of *training*, rather than algorithmically, and that the practices-or-abilities that implement algorithmic elaboration are neither necessary nor sufficient for this sort of practical elaboration. Besides this negative characterization, what can we say positively about what training is? Most generally, I think of training as a course of *experience*, in Hegel's and Dewey's sense (processual, developmental *Erfahrung* rather than episodic, self-intimating *Erlebnis*) of a feedback loop of perception, responsive performance, and perception of the results of the performance. When we think about the practices-or-abilities that *implement* elaboration-by-training, we can think about them on the side both of the trainer and of the trainee (though both learning—training without a trainer—and self-training, which is not the same thing, are also important species). A course of training implements a *pedagogical* elaboration of one set of abilities into another. We can think of it very abstractly as having as its basic unit a stimulus (perhaps provided by the trainer), a response on the part of the trainee, a response by the trainer to that response, and a response to that response by the trainee that involves altering his dispositions to respond to future stimuli. A constellation of such units constitutes a *course* of *training*.

6 Algorithmic pedagogical decomposition and pedagogical politics

I am suggesting that what, in a course of training, is most analogous to algorithmic elaboration of abilities is *pedagogical* elaboration in the form of a training regimen. In rare but important cases in early education, we have *completely solved* the problem of how to pedagogically elaborate one set of abilities into another. What it means to have a *solved pedagogical problem* for a population with respect to an output practice-or-ability is to have an *empirically sufficient conditional branched training regimen* for it. This is something that as a matter of contingent fact can take any novice from the population who has mastered the relevant range of primitive practical capacities, and by an algorithmically specifiable Test-Operate-Test-Exit (TOTE) cycle of responses to her responses *in fact* (though without the guarantee of any principle), get her to catch on to the target ability. For us, training pupils who can already *count* to be able to *add* is essentially a solved pedagogical problem in this sense. That is, starting with pupils of widely varying abilities and prior experiences, who share only the prior ability to count, there is a flowchart of differentially elicited instructions, tests, and exercises that will lead all of them to the target skill of being able correctly to add pairs of arbitrary multi-digit numbers. A common initial lesson or exercise is followed by a diagnostic test. The results of that test then determine, for each pupil, which of an array of possible second lessons or exercises is appropriate, followed by further tests whose results are interpreted as differentially calling for different exercises, and so on. This flowchart determines a TOTE cycle of training that incorporates a *pedagogical* (as opposed to an executive) *algorithm*.

I am told by those who know about these things that teaching *multiplication* to pupils who can add and subtract is also, in this sense, a completely solved pedagogical problem, but that in spite of massive investigative efforts to date, *subtraction* remains an essentially *un*solved pedagogical problem, and *division*, in the form of mastery of fractions, a tantalizing, so far intractable pedagogical *mystery*. In the absence of a complete practical pedagogical algorithm, those charged with eliciting and developing such skills must fall back on rougher heuristics and the sort of practical know-how gleaned from many years of trial-and-error training of a wide variety of candidates.

Incompletely solved pedagogical problems—not just in specialized cases in elementary education, but at all levels and across the board—raise a broad issue of institutional *politics* that seems to me to penetrate deeply into our understanding of, and attitudes towards, the society as a whole. Augustine marveled at (and rode three days on a mule to test) the rumored ability of a monk to gather the sense of a text without pronouncing aloud the words on the page and then listening to them. One of Samuel Pepys's distinctive qualifications for his position as Secretary of the Admiralty was his mastery of the arithmetic required for double-entry bookkeeping. Today we take it for granted that we can train almost everyone to read silently and to add up long columns of figures. But for abilities for which the pedagogical problem has not been completely solved, where we do not yet have an algorithmic decomposition of the practical training process, candidates who exhibit all the relevant primitive abilities are *de facto sorted* by the training regimens we do have, not only by the number of iterations of the TOTE loop it takes for them to acquire the target ability, but also by whether they can be brought to that point at all. Matter-of-factual

PP-necessity relations among practices-or-abilities require that the outputs of some training regimens serve as the inputs to others—that some of the abilities treated as primitive (as practically a priori) by the one are achievable only as the target abilities (the practical a posteriori) of others. It follows that the effects of failing to acquire one ability—falling into the missing, incompletely mapped portion of an ideally complete pedagogical solution of which some actual training regimen is a mere fragment—will be strongly cumulative within a sequence of courses of learning-and-training.

The broadly political issue I want to point to concerns how, in the context of these very general considerations, we should think about one element of just treatment of individuals by institutions. We might, as a demand of justice, or simply as a counsel of social engineering, want some kinds of rewards to be proportioned to productive achievements, according to some definition of the latter. Among the crucial necessary conditions of any such achievement is the possession of certain skills or abilities. It seems that there are two basic attitudes (defining a spectrum between them) that one might have toward any target ability for which we do *not* have a pedagogical algorithm codifying a complete solution to the training problem.

One attitude is that it is just a brute empirical fact that people not only have different abilities, but are in important respects more or less able. With respect to any sort of target ability, some are more trainable, better learners, than others. What is being assessed here is the practical-elaborative abilities that *implement* the PP-sufficiency of some set of primitive abilities for the target ability, in the context of the course of experience yielded by a training regimen. The training regimen not only inculcates or elicits the skill that is its target, but along the way sorts candidates into those who can, and those who cannot learn or be trained in it, as well as into those who learn it faster or more easily—measured by how long or how many steps it takes to get them through the pedagogical flowchart to the exit of that practical labyrinth. On this view, it is *compatible* with just dealing, and perhaps even *constitutive* of a dimension of justice, for an institution to factor this sort of second-order ability into its reward structure.

The view that forms the opposite pole of the dimension I am pointing to focuses on the relativity of the hierarchical sorting of candidates into more or less trainable to the training regimens that happen to be available. Different regimens might produce quite different rankings. If that is so, and the fact that we have actually implemented one set of training procedures rather than another is quite contingent, conditioned by adventitious, in principle parochial, features of the actual history of the training institutions, the experience and skills the available trainers happen to have, and so on, then the inferences from the actual outcomes of training either to the attribution of some kind of *general* second-order ability or to the *justice* of rewarding the *particular* sort of second-order ability that really *is* evidenced thereby—just being more trainable, or more easily trainable, by the methods we happen to have in place and apply—are undercut. Our failure to provide a more comprehensive set of training alternatives, to have filled in the pedagogical flowchart more fully, ultimately, to have completely solved the relevant training problem, should be held responsible for the training outcome, rather than supposing that sub-optimal outcomes reveal evaluatively significant deficiencies on the part of the trainee. At the limit, this attitude consists in a *cognitive* commitment to the effect that there is, in principle, a complete pedagogical algorithmic solution for every target skill or ability to the possession of which it is just to apportion rewards, and a *practical* commitment to find and implement those solutions. It is an extreme, indeed utopian, pedagogical egalitarianism.

Taken to the limit, the pedagogical egalitarian view may seem to rest on a literally unbelievable premise: that whatever *some* human can (learn or be trained to) do, *any* human can (learn or be trained to) do. And the evaluative component implicit in the cognitive commitment as I stated may seem no more plausible: that there is something wrong with rewarding any ability of which that claim is *not* true. A more defensible version of pedagogical egalitarianism results if the latter commitment is softened so as to claim merely that special arguments must be given in each case for the valorization of differences in ability for which we have found no complete pedagogical solution. The first element of the cognitive component can correspondingly be interpreted in terms of the practical commitment, so as to claim that we have no right to assume, for any given skill or ability for which we have as yet no complete pedagogical solution, that that is because there *is*, in principle, no such solution.

The empirical-hierarchical attitude is *conservative* in treating extant institutionalized training regimens as given and fixed, and the utopian pedagogical egalitarian attitude is *progressive* in its commitment to transform them. As political attitudes, they articulate one dimension of the nature/nurture aspect of traditional right/left alignments. Between them lies a whole array of more nuanced principles for assigning reciprocal, co-ordinate responsibility to training or trainers, on the one hand, and trainees on the other. We need not simply choose between the strategies of holding actual training regimens fixed and hierarchically sorting humans with respect to them, on the one hand, and holding the actual practical-elaborative abilities of humans fixed and sorting training regimens with respect to them, on the other.

But my purpose in gesturing at this issue of pedagogical politics here has not been to recommend one or another way of approaching it. Assessing the plausibility of the broadened, practical version of the thesis of artificial intelligence led to the notion of practical PP-sufficiency by training. My aim in this final section has been to lay alongside the postulate of universal practical *executive* algorithmic decomposability of discursive abilities, characteristic of AI-functionalism, the postulate of universal practical *pedagogical* algorithmic decomposability of discursive abilities characteristic of utopian pedagogical egalitarianism, and to point to an issue of considerable philosophical, cultural, and political significance that it raises. As a result, the argument of the lecture as a whole has described a narrative arc taking us from Turing, through Wittgenstein, to Dewey.

My last three lectures will address *modal* vocabulary, *normative* vocabulary, and the pragmatically mediated semantic relations they stand in to ordinary *objective, empirical*, and *naturalistic* vocabularies, and to each other. I will argue that both the *deontic* vocabulary of conceptual norms and the *alethic* vocabulary of laws and possibilities can be elaborated from and are explicative of features necessarily exhibited by any autonomous discursive practice. Thinking about the pragmatically mediated semantic relations they stand in to each other turns out to provide a new way of understanding the *subjective* and *objective* poles of the intentional nexus of knowers-and-agents with their world. Along the way, I will show how *normative* vocabulary can serve as an expressively bootstrapping pragmatic metavocabulary for *modal* vocabulary, and, in the fifth lecture, how that fact makes possible a new sort of *formal semantics* for logical and modal vocabulary, as well as for ordinary empirical descriptive vocabulary.