
You've worked hard. Time for recreation. Youthful metaphysicians, rejoice! For we're about to investigate a few key concepts in revisionary metaphysics, specifically in the naturalization project. Naturalization, you'll recall from the Introduction, is the project of showing that the properties of the mind, such as intentionality and consciousness, are fully explainable in terms of the physical properties of physical things. It's time for folk psychology to be judged to be either a partner in this project or an obstacle to it. My own view is that it needs to be supplemented and revised in accord with contemporary science. I say supplemented and revised rather than reduced or replaced—something naturalizers more radical than I am would say—because the ontology of folk psychology is sound. There really are the representational states which folk psychology is all about. In the ideally coherent modern worldview, which we metaphysicians seek, folk psychology survives largely intact.

Some naturalists may say that we've already departed irreconcilably from naturalism, for they define naturalism more severely than we did back in the introductions to the book and to Part IV. They say that naturalism rejects any claim to knowledge that's not a part of science. We only said that naturalism considers scientific knowledge to be the gold standard of knowledge, not that it constituted all the knowledge there is. So our acknowledging a peculiar form of knowledge in indexical beliefs (sections M and N), and later considering empathy to be an indispensable form of knowledge (sections T and U), not to mention our dalliance with knowledge of values and god (section O)—all that might seem to put us beyond the pale.

To be quite fair to ourselves, we only acknowledged forms of beliefs, not of knowledge. But I do go further and admit that there are indexical knowledge and empathic knowledge, and they're extra-scientific, a claim I'll defend in section Z§4. I don't think the admission puts me beyond the pale of naturalism, however, for my adherence to science's being the gold standard of knowledge means that I believe scientific knowledge is the sort of knowledge that discovers ontology, and that's naturalism enough. The indexical knowledge we've admitted imports no new ontology to explain its referents. As we saw in section N§4, indexical knowledge is a form of
knowledge that can be fully understood without requiring an ontology additional to that recognized by science. The empathic knowledge does import the ontology of folk psychology, namely intentional states of animals, but science will grow to accept it, or so I claim.

§1. Two kinds of naturalization, coming from two kinds of explanation.

First, let's introduce a distinction that evolutionary biologists find useful, namely, the distinction between proximate and ultimate causes, mechanisms, and explanations. The distinction was introduced in 1938 by J. R. Baker and popularized by Ernst Mayr. Mayr suggested that if you're asking how something occurs, you want a proximate cause of it. If you're asking why it occurs, or how come it occurs, you want an ultimate cause. Proximate causes of biological features are the physical and chemical mechanisms that underlie them. Their ultimate causes are historical, why they came to be in the first place. The proximate causes are studied in functional biology; the ultimate causes in evolutionary biology. This same distinction divides two ways of conducting the project of naturalizing the mind.

The distinction between two kinds of explanation carries over into the naturalization project, yielding two naturalization projects: Proximate naturalization shows the mind to be a physical and chemical system, and intentionality to be a physical and chemical manifestation. Ultimate naturalization shows, by reference to evolution, that the mind is as continuous with the rest of nature as is life itself.

The word “why” suggests reasons as well as causes, but the intent here is only to give causes. People often describe a trait’s “reason for being” in terms of its supposed purposes for existing, for instance, its contributions to the reproductive fitness of an organism possessing the trait. Purpose is something minds give to their creations, however, and so belongs among the things the naturalizers must explain, and not among the things assumed by their explanations. Ultimate naturalizers must avoid the pitfall of explanation by purpose. Explanations in terms of natural selection causing a lineage of organisms to become more adapted to its environment are not explanations in terms of purpose. For example, people sweat, and there's a proximate story of how they do and another story of why ultimately they do. According to the latter story, it happens that sweating's been adaptive enough to make the organisms that sweated able to leave more descendents, and so natural selection operating on their lineage caused the gene for sweat glands to become prevalent in contemporary members of the lineage. Please satisfy yourself that no reference to purpose occurs in that story of
adaptation. Not all stories of origins are stories of adaptation however. The ultimate cause of sickle cell anemia, which is a genetic disease, is a tale of the utter mindlessness of the evolutionary process. Natural selection worked in both cases, so it's hardly a teleological factor in evolution.

Naturalizers sometimes develop a predilection for one of the projects and a distaste for the other. They are either proximate naturalizers, disdaining the ultimate naturalizers, or vice versa. I, syncretist that I am, do not share this prejudice. Nevertheless, we must note something like a civil war going on among the naturalizers.

Some naturalizers are simply interested in the physical and computer-like underpinnings of intentionality. They're interested in proximate causes. It may be that proximate naturalization is the more revisionary of the projects, the less consistent with folk psychology. Thus it has the attractiveness of scandal. Nevertheless, it's just as interesting to show how an incremental evolution of intentionality can be true in a nonplatitudinous way, and how a belief in its evolution is consistent with folk psychology. Some of the proximate naturalizers are exclusively interested in the proximate causes, because they don't see anything but vapid truisms and untestable conjectures coming from the search for the ultimate causes of intentionality. They believe in the evolution of the mind, or at least of the brain, but, ho hum, that's all you can say. So, as one of these proximate naturalizers (Fodor) says, "So, please, spare me; no Darwin."2 If it comes down to vapid truisms and untestable conjectures, however, both types of naturalizers exhibit plenty of that.

There's another motive for suspicion of the evolutionary naturalizer. The belief is very common today that with the coming of mankind one must recognize a jump to intentionality, which is next in magnitude only to the jump from nonlife to life. That is, both jumps are so great that, as the earlier jump went from the realm of chemistry into the realm of biology, the later jump went from biology into a third realm. Mind escapes biology and belongs to psychology, sociology, and anthropology. The proximate naturalizers take this to be, oh, so obvious that they dismiss anyone who looks for small incremental steps of intentionality connecting humanity to the rest of the animal kingdom. The fact that the brain is obviously a biological organ does not allay their impatience. For even if there are only small differences between our brains and, say, the chimpanzee's, and we can trace both species back by small decrements to an extinct common ancestor, nevertheless those small differences in brains lead to huge differences in mentality. It's a fallacy to argue from a small difference in brain or genes to a small difference in mind. So Fodor says.

We must grant the fallacious nature of the inference, but not grant that
evolutionary naturalizers commit it. My own interest in the naturalization project stems mostly from a conviction that human intentionality is not so big a leap within the animal world as these proximate naturalizers think. I believe this can be proved directly from an examination of mentality, without resorting to inferences from brain studies. Mentality itself is something that evolved in small increments throughout evolution. Not all of the increments evolved as adaptations. I’m convinced that all of it, including human intentionality, is very much a part of biology. Nevertheless, given the hostile intransigence the evolutionary naturalizers meet, they cannot even count on the support of the proximate naturalizers. It’s not reciprocated; ultimate naturalizers don’t have the same animus toward the proximate naturalizers as many of the latter have toward the evolutionary naturalizers, although I, for one, do yawn when the proximate naturalizers tell me what consciousness is.

Ultimate naturalizers often deserve the scorn directed toward them for their uncritical appeals to adaptation. As the example of sickle cell anemia shows, not everything that evolved is adaptive. Define evolution so that this can make sense, as the change in the frequency of genes in a population of organisms from one generation to the next. The causes of evolution are whatever causes changes in gene frequencies. Thus mutation is a cause, as is sexual reproduction, any act of which discards half the genes of the parents. So there’s nothing in the definition of evolution that necessitates the changes being adaptive. Natural selection is the greater or lesser contribution by individuals of their genes to the pool of genes in the population's next generation, as a result of the environment facilitating some of the individuals and hindering others. Natural selection frequently prevents evolution when the environment is stable, but when a population is not adapted to its environment, only natural selection can lead to organisms with traits that help them survive and reproduce in their environment (i.e., adapt). The lazy idea of the job of the ultimate naturalizer of the mind is that it's simply to assert the adaptiveness of every mental trait, and presume natural selection formed it. Would that the project were that easy!

Let’s put aside the internecine disputes: There are two naturalization projects, which should fit together the way functional and evolutionary biology fit together. There's proximate naturalization, and there's ultimate naturalization. Those two revisionary metaphysics can be pursued together, along with the descriptive metaphysics of folk psychology, without those interested in one of them being snooty about the pursuits of the others. Some of us pursue all three.
§2. The mechanism of proximate naturalization: negative feedback.

There are two great successes of the proximate naturalizers, both stemming from discoveries made in the nineteenth century, but which the pressures of world war II forced into our technology. They're the negative feedback mechanism and computers. If you understand the former, you know how to naturalize purposiveness and meaning. If you understand the latter, you know how to naturalize rationality.

First, let's focus on the fundamental mechanism that captures within a naturalistic framework all purposefulness and meaningfulness. It's the negative feedback mechanism. As a mechanism it does something, but it's mechanical in a special way: It collects information about its own activities—that's feedback. It compares the feedback with a representation of the goal of its activities. The comparison is just the finding of a difference between the two by a process of subtraction—that's the negative. If the system is not yet successful, i.e., the result of the subtraction is not zero, it corrects its activities in directions which promote achieving the goal, i.e., getting a comparison result equal to zero. The control loop of the negative feedback mechanism is the fundamental principle of robotics.

That description employs mentalistic terminology to describe the machine: “collect information,” “goal,” “success,” and so on. The underlying process is transparently purely physical, and we can eliminate all these terms from our descriptions, as I'll now show with a mini-course for philosophers about engineering. Demonstrating their eliminability also shows how they may be reintroduced by definition, if it were useful to do so, and indeed it is useful. We're giving a naturalistic analysis of the basic mentalistic notion, what it is to make a representation, what it is to be a made representation, not just a found one. The representations figure in contentful states such as intentions. Given a system sufficiently complex to warrant a mentalistic description of it, we can state a necessary condition for its state to be an intention-to-act, as distinct from an intention-to-refrain. If a state is an intention-to-act, then it establishes and activates a negative feedback mechanism within the system. The content of the system's intention, what it intends to do, is the foresign in it, about which more later. I only say this is a necessary condition for being an intention-to-act, not a full analysis, because obviously I've left out things like consciousness, which many think are also necessary. Nonetheless, the machine's information states are the physical symptoms of the direct attributions and the holophrastic utterances discussed in sections N and U.

Not just any stabilizing mechanism will do as the generator of all purposefulness and meaningfulness, even if it's come to be called a negative
feedback system. We must distinguish systems that merely tend toward an equilibrium from those systems that are true negative feedback systems. There are many kinds of systems that can return to equilibrium from an arbitrarily large range of different initial conditions. And many systems reach equilibrium by having some links of the system compensate for other links that have been disturbed from their equilibrium values. The compensations work to restore all states of the system to their equilibrium values. That's not enough to make a system one that operates by feedback, however. For example, the geyser in Yellowstone National Park, Old Faithful, is a mechanism reestablishing its equilibrium at regular intervals by sending a stream of water and steam a hundred meters into the air. It does not operate by comparing feedback to a sign of a goal state. Another example: Processes that affect the reflectivity of clouds may stabilize atmospheric temperatures. When ocean surface temperatures rise and micro-organisms emit more aerosol particles into the atmosphere, clouds become whiter and reduce sunlight's warming of the atmosphere. The surface temperatures of the ocean will tend to stability. This so-called CLAW cycle does not respond to information, despite its having a counter-active link in the cycle. The albedo of clouds is not signaling the oceanic microorganisms to modulate their activities for the sake of the global ecosystem. No sky-high cirrus transducer is preserving an isomorphism between the input and output energies, analogous to sensing. Many systems in nature are just energy merry-go-rounds, with some steps that dampen the amplification of effects. For example, populations of predators and prey may constitute a cycle of predation and evasion that produces a stable population. The interlocking of many such cycles forms what we call the balance of nature. I'm not talking about any of that, when I talk about negative feedback systems. On the other hand, I must distinguish the two kinds of systems in nonmentalist terms if there's to be any hope of a naturalistic capture of a contentful state.

Three naturalistic and non-question-begging things distinguish true negative feedback systems from systems that merely tend to equilibrium. First, they compensate for disturbances of their states by having a causal path that's spatially distinct from the main path. Old Faithful has no such independent loop. In genuine negative feedback systems, besides the main path where the input of energy is high and most of the work is done, there must also be distinct low energy paths. The low energy channels often begin at transducers and terminate in unusual arrangements of matter that trigger but don't power the high energy channels. Such a peculiar dimorphism cries out for explanation, which comes in the form of a semiotic description of the same configuration. So, secondly, information collects
at the sensors and flows along these low energy paths to comparators where it's exploited for the control of purposive behavior. A path that transports information is not like a path for transporting a material such as gas or hot air. Systems that exploit information through channels dedicated to relaying it are more robust in the face of disturbances and respond with more speed and discrimination. Thirdly, negative feedback systems need not have as goals their states of equilibrium as a physicist understands equilibrium, namely, thermodynamic equilibrium. The physicist will tell you that the equilibrium point of the human body is attained after death. The goal of a human being is not that. So the goal of a negative feedback system is some stable condition distinct from its physical equilibrium point. One cannot infer that, because a system stabilizes at a condition distinct from physical equilibrium, it also fulfills the first two conditions. Many systems fulfill the third condition without fulfilling the first two, for instance, the sun with its ample supply of fuel for fusion. The other two conditions are independent diagnosticators of genuine feedback systems, more sensitive than the third condition is.

**Information** must be distinguished from the energy that carries it. “The news” has at least three features that distinguish it from energy. As the ancient Greeks said, light that's reflected from the surface of an object carries away the *form* of that object, that is, its shape, size, texture, and color. More exactly, the light carries with it the mathematically describable structure of the surface. The forms in energy like light are information or, if you prefer, potential information. The information in the energy is therefore a relative property, the form of some object, but the energy carrying the form is not a relative thing; it has its existence in its own right. Secondly, energy is a conserved quantity, but information is not; at least not as ordinarily understood.\(^5\) It can be destroyed. And whereas energy is degraded in work and becomes useless, information does not have to disappear when it is used; that is, it need not be used up. It can be duplicated without destroying or depleting the information. (I assume that the entropy added by the use of information can be dumped, so that the information is kept after being used even if the totality of information in the universe is depleted by its use.) Thirdly, information is a macro-pattern distributed over energy. The whole pattern is not derived by adding up its parts, any more than the information in a paragraph is merely the sum of the information in each of its sentences or words or letters in the words. Suppose that the information in “all men are mortal” is the set of all the sentences it implies. Now create the union of that set with the set of all the sentences implied by “Socrates is a man.” The union does not contain the sentence, “Socrates is mortal.” There's more information in the two
sentences together than in the sum of the information in each separately.

A negative feedback system exploits the information that comes its way as signs or representations. Its "setting" is its foresign, its sign of a possible state of the system, which because of the way the sign functions in the system we may identify as its goal state. A foresign's mood is imperative. The feedback is a sign too, a sign of conditions as they allegedly were when measured. It's declarative in mood. Feedback signs are good examples of the representations we discussed in section P. We noted there that some signs are simply found in nature, and organisms exploit them. In appealing to found signs, therefore, we've not lapsed carelessly into using the mentalistic concepts we wish to explain. In section Q's terms they may be medadic twiddlers, which make up the subsymbolic sign system we recognized in section R.

The feedback, traveling along paths in the system, comes into proximity to the foresign, and the two get "compared": signs of conditions as they allegedly just were with signs of conditions as they allegedly should be. If we're not to beg any questions about the analysis of mentalistic notions, that description should be replaced by a physical description of the comparison process. Later I will show how a bimetal does it in a thermostat. In simple systems the signs are quantitative measures, and the comparison is like subtracting the feedback from the foresign. Any result other than zero triggers a correction, bringing results of later comparison closer to zero. Zero marks the end attained. These systems are self-governing, which is why the science of them is called cybernetics. If "governetic" were a word, we would see in it the Greek "cybernetic" (where "c" transliterates a Greek sound like the "c" in "actor"). Let's apply these ideas to figure Y-1.

Four types of components occur in the control diagram:

The Dot = measurer: measures, e.g., temperature, and relays (i.e., feeds back) the measurement to comparators.

The Circle = comparator: subtracts any feedback from a setting (foresign) or from another measurement, or it makes a more complex comparison than subtraction, and produces an error signal.

The Square = operator: responds to non-zero difference signals in a way that makes later error signals tend toward zero.

The Broken Line = system's boundaries. Measurement can occur inside or on the exterior surface of a boundary.
Assume the control diagram is of a proportional regulator of temperature in a house. I'll describe its operation in terms of its signifying goals and existing conditions and its interpreting its signs. (Reminder: We do not interpret “signifying” and “interpreting” as mental or communicative.) The rightmost measurer measures the temperature of the air being pushed into the house from the furnace/cooler blowers. This measurement is relayed along feedback loops to the leftmost comparator and is subtracted there from the foresign of the desired temperature. The resulting error signal, if not zero, ultimately changes the action of the second operator, which is the furnace/cooler. If the difference turns positive, the furnace's activity is increased; if negative, the cooler's. Before the instruction is carried out it is further modulated:

- The first operator responds to the error signal by resetting the middle comparator to a sub-foresign. This signifies the amount of change in temperature that the system should make.
- The bottom measurer measures the temperature outside the house (more strictly, its own temperature at its surface) and relays it to the bottommost comparator, which compares it to the interior temperature. (No foresign or output to an operator here.)
- The middle comparator compares the difference, just described, to the sub-foresign by more than mere subtraction. In effect it substitutes a new error signal for the original one to account for the rate of gain or loss of heat through the walls of the house.

The final operator is the furnace/cooler. Its level of activity adjusts so that
the temperature fore-signified at the original comparator will be reached.

For example: Suppose the system’s temperature is 15°C. The measurement goes to the comparator. It computes the difference between this measurement and the setting. Suppose it’s set for 20°C. The difference is the number 5, called the error, $e$. The operator then receives the error signal and responds by directing a flow of gas to the furnace at a rate that will reduce the number $e$ to zero. When $e = 0$, that’s the end realized. The system then ceases changing the means; the rate of flow of the gas to the furnace stays constant.

Although the foresign of the goal is on the control diagram and the means to reaching it are there also, the end is not shown on it, for the end is not a cause of its own coming to be. The end is the temperature of the house at which the error signal equals zero. To show the end attained we need a time-flow diagram. We’ll ignore the complicating effect of the cue stimulus and the first operator in figure Y-1, to emphasize the key concepts:

Figure Y-2. The response over time of a negative feedback system (simpler than the one in figure 1) to error signals. As time passes, the negative feedback system tends to reach the condition identified by the zero error signal and, once the system attains it, thereafter acts to maintain it.

We can write equations that describe the action of a negative feedback system. Equations can be written in various forms. For example, a linear equation may be written in slope-intercept form, or in point-slope form, or in other forms. Here’s a form of linear equation that is especially important for the point we’re making. Call it the goal-minus-feedback form:

$$r = \frac{1}{t}(g - f)$$

$r$ = the adjusted rate of action; $t$ = the time needed to achieve the goal at the
The formula approximates the operation of feedback in our continuously adjusting thermostat-furnace system. The thermostat is set at 20°C. So 20°C is the goal, but it's winter, and the door's been opened and the house temperature has dropped to 15°C. So the feedback is 15°C. This difference of 5°C is the error, \( e \), as I said. The “action” of this system is the burning of gas to heat the air in the house. The initial rate at which the gas was being burned was sufficient to raise the house temperature to 20°C in two minutes; let's say. The reciprocal of this number, \( 1/(2 \text{ min.}) \), is constant in all our calculations. We can now put numbers for the variables. 

\[
r = \frac{1}{2 \text{ min.}}(20^\circ - 15^\circ) = 2.5^\circ \text{ per minute}
\]

Gas will now be burned at a rate that would increase the temperature to the desired level in two minutes. (Notice in figure Y-2 that each straight line segment, if doubled in length, would intersect the 20°C line.) We don't do that, since it would lead to overshoots and compensatory corrections that would cause a perpetual oscillation of the temperature. In fact, the burning of the gas increases the temperature to 17.5°C in one minute. Let's imagine that the operator adjusts the burning once every minute instead of continuously, so that we can avoid calculus. After one minute the rate of action adjusts again:

\[
r = \frac{1}{2 \text{ min.}}(20^\circ - 17.5^\circ) = 1.25^\circ \text{ per minute}
\]

After the second minute the temperature has risen to 18.75°C and the rate of burning gas adjusts downward again:

\[
r = \frac{1}{2 \text{ min.}}(20^\circ - 18.75^\circ) = 0.625^\circ \text{ per minute}
\]

So after the third minute, the temperature has risen to 19.375°C. As the feedback temperature approaches the goal of 20°C, the rate of action slows so that the feedback temperature approaches the goal only asymptotically. When the thermostat can detect no difference between goal and feedback, \( e = 0 \) and the goal is reached for all practical purposes.

Recall from section B§1 the reciprocity of a desire's intensity for a goal and a belief's degree of certainty that it's been achieved. The former decreases as the latter increases, according to decision theory. Figure Y-2 shows an analogous reciprocity between the amount of the feedback's negativity and the action's intensity.

I mentioned earlier that I'd cash out the mentalistic connotations of the idea of comparison and subtraction, which we used to describe the feedback system's operations. To understand how the subtraction is done physically, we need the concept of a \textit{bimetal}. A bimetal is a strip of one metal, say brass, bonded to another strip of metal, say iron. Since the different metals contract and expand differently in response to differences in temperature, a bimetal will curve in response to changes in temperature.
as one side of it expands more than the other side. Comparison of feedback to goal is the orientation of a bimetal within a circuit which the bimetal can open or close—it’s opening’s the goal—and the degree of its bending in response to the attained temperature of the ambient air—that’s the feedback. The bimetal’s initial orientation determines how far it must bend to close the circuit—that’s the comparison.

Figure Y-3. A bimetal. Upper figure: Two states of a bimetal. The lower state is the effect of heating. Lower figure: A typical way a temperature is compared to a goal for the temperature is to insert a bimetal into a circuit in such a way that it closes the circuit if the temperature is too low, which the bimetal detects because bimetals bend and unbend in response to temperature. The dial at D controls the orientation of the bimetal at E.

The goal-minus-feedback formula describes the behavior of the bimetal in the circuit and the operator’s response to its opening and closing the circuit. I’d like, however, to say that here we have the toehold for a mentalistic description, as if it were the manual for translating this system’s signs. You may doubt that the goal-minus-feedback form of the equation should be interpreted as a manual for the translation of nature’s own signs into human language. For even systems that are not negative feedback systems can be described by equations in so-called goal-minus-feedback form, systems like pendulums and populations of animals. Surely, a pendulum does not have a foresign of the position at which it comes to rest, as though that were its goal. How then do we know that the equation is a schema for translation for negative feedback systems, but not for pendulums? My reply is that, when the other distinguishing features of engineered negative feedback systems are in place, the three we mentioned earlier, the mentalistic reading
Part IV, Section Y

511

of the formula begins to make sense. All that we need add is complexity.

Of course, the artifact of a furnace/cooler system controlled by a thermostat, which we used to introduce the concepts appropriate for negative feedback systems, is too simple to be like a living thing, but the needed complexities are easy to imagine. First, in place of each measuring device we put millions of them, whose measurements can be thought of as a long vector input to a parallel distributed network of the connectionist sort. A human eye, for instance, is an organ with millions of sensors all feeding data to the brain. The connectionist network that's in the brain's optical cortex is sophisticated enough to analyze and classify its input. Its output is the signal we've called the feedback or a cue stimulus. Secondly, the output of the operator need not be the single discharge of energy, which the single arrow titled “means to goal state” seems to portray. Instead it may feed into another connectionist network that operationalizes the command and activates millions of muscle fibers coordinately. Finally, many negative feedback systems can be organized hierarchically so that there are goals, sub-goals, sub-sub-goals, and so on. They interlock for mutual protection and for increased range of adaptive response.

This complexity sinks any defining of sign-types by their functions; functions won't individuate them. In the last subsection we'll set in place the conceptual means to treat signs as definitionally prior to functions.

A sharp critic will challenge us, “Ok, supposing the negative feedback system does ‘intend’ to heat up the place, why does it intend to? What are its reasons?” We might concede that we've so far captured only part of the folk psychological theory of purposefulness, especially as it applies to human beings. But the theory applies to simpler beings also, for we may opt out of answering such questions about a spider's spinning a web, for instance, when we reply, “No reason; it's instinct for survival and reproduction.” We cannot say even that much about the human negative feedback system, however, since it does have many goals not subordinated to an overarching goal of fitness, in contradistinction to all other living things. Nevertheless, it does contain the primordial contentful states. Recall the boiler room analogy from sections N§5 and P§7. There we looked into the room from above; here we've been inside inspecting the machinery. As we huff and puff our way back up, we note the long flight of stairs ahead of us, representing the continuities and growing complexities for the science of the distinctiveness of the human mind to explore. Let's resist the incessant calls of the antinaturalists to find just one major dichotomy splitting us off from all else. By the time we reach the last step up to the captain's quarters, much of the human mind will be already in place.
§3. Intentionality comes from the mind’s being interconnected computers.

At the other extreme of the continuum of life, we do want to understand naturalistically what it is for one to embed one's intention in a framework of reasons. We resort again to complexity for the account we seek. The mind computes, and the more rational we are the more resources of computation we make use of. This idea is old hat today. I'll not belabor the analogy of mind to computer. The proximate naturalizer's project is to show that the mind achieves the rationality component of its intentionality by operating as a computer.

Or rather the brain is like a beehive or ant nest, where each neuron is one bee or one ant. The individual bee and ant are computers, and each individual neuron is also. Bees communicate with each other, and so do ants, but their communications are not expressions of an inner meaningfulness. It's as if a bee's recall of where the nectar can be found is there for the first time in its dance in the hive, from which other bees learn where to collect more of it. The conceptual roles of the components of the dance are their social roles only. The recall is not first in its cephalic ganglion (proto-brain) in beementalese. The same emptiness or pure exteriority characterizes ant communications. Behaviorist theories of meaning say all there is to say that's true of these systems. Just as it's only by dancing that the bee recalls the nectar places for other bees, so likewise each neuron in your brain talks to the other neurons, but does not express itself to them.

Now suppose that ten billion bees or ants get locked into fixed positions inside a robot and must make do with directing the robot to do all their work for them. The chat among the ants continues, but what had been a communication system without expressing anything inner finds itself on the inside. That which has no inner is now the inner. No bee or ant leaves this robot; instead it commands the robot to do what it would have done if it had been able to leave. When the ant population's robot—we can call her Aunt Hillary—meets a robot belonging to another ten billion ants in a similar fix, it's expedient for the two of them to develop another communication system expressive of the inner in each robot, so that they may cooperate. This communication system looks the very opposite of emptiness and pure exteriority. To collect the payoff of this analogy, convert all that talk of ants into talk of neurons, and you have a conception of human intentionality. In the Introduction I listed seven rungs on the ladder of increasingly complex aboutness, with mind or intentionality beginning at rung three. Rung three was aboutness interiorized. The analogy we have just rehearsed helps us understand how interiority evolved.
Of course, analogy has its limits. If the human brain's neurons are imitating whole bees, what do the neurons in the bees do? Well some of their neurons and some in human beings just implement the negative feedback systems which give us the first rung of the ladder, purposive exploitation of information in nature. So much excellent material already exists on this subject that here I only wish to examine the most vivid attack on the analogy and show why it's inconclusive. In section §§4-6 I disengaged a form of conceptual role semantics and state-space semantics from the naturalization project, just as I did with respect to sententialism in section P, so that they might find a home in folk psychology. I did mention that conceptual roles and state-spaces are often ingredients in naturalization projects. The goal of the naturalizing philosophers is to show that the states of brains and computers might play meaning-giving roles without defining the roles in terms that assume purpose or logic. We were not so confined when we were doing folk psychology. It's hard to delimit meaning-giving roles without making use of rationality and purpose. Perhaps there are naturalistic substitutes. Perhaps biological fitness will serve as goal, and the finding and using of means to reach that goal are rationality. The greater the stretch and convolution of connection between stimulation and the responses it prompts, the greater the rationality of the responder. If so, then the proximate naturalizers' state-space semantics and conceptual role semantics can simply be called causal role semantics, because they are reductionists: The reduction of semantics to causal roles is explicitly a part of the naturalization project as they develop it.

This is not just armchair programmatic. Douglas Lenat's company, Cycorp has developed computer programs that model conceptual roles. The public versions, available from the company's website, contain about 47,000 concepts whose interconnections can be explored. The company's website also displays propositions that the cyc program "knows." It's unlikely you'll ever see the same proposition displayed twice.

Let's consider an objection to this form of naturalization, which goes so well with sententialism and conceptual role semantics. A thought experiment, part of what is called the Chinese room argument, tries to break the connection the naturalizers draw between our understanding of meanings and our calculating in LOT, leading to appropriate, purposive action. The issue is whether thinking in LOT, which involves understanding, just is the naturalistic causal process of symbols being manipulated by a program of the sort that runs computers. That would be the proximate naturalizers' reduction of folk psychology's sententialism and conceptual role semantics. The experiment in my words:

Imagine a computer that manipulates symbols in such a way as to
simulate the results of understanding, namely, it gives intelligent responses. It outputs in Chinese appropriate answers to questions formulated in Chinese. Does it understand the Chinese inputs and outputs?

You may not intuit that it need not understand the questions or its answers by virtue of what it does. To see that it need not, we continue:

Analogously, if you just imagine yourself duplicating the symbol manipulation processes by which it arrived at the answers, you intuit that you don't. All you do is follow the rules of the computer program. You accept Chinese inputs, follow the program's rules for manipulating them, and send out the Chinese outputs. The people fluent in Chinese who gave you the questions and accept the answers may be amazed at how clever you seem to be. But you know you don't understand the questions or answers. And neither did the computer!

You may understand when you think you don't, and you may think you understand when you don't. It's clear, however, that here what you think is so is so. Applying the result to the naturalizers' thesis that the mind is a computer, if thinking in LOT were just the manipulating of sentences according to a program, it would not include understanding meanings.

I criticized this argument in a note to section R on connectionism for its targeting a narrow thesis about what understanding consists of and then pretending its target is broader. We’ve found much to understanding that a narrowly computational model of it leaves out. Nevertheless, I’ll just add here a defense of the naturalizing sententialists. The thought experiment is weak even against them in that it trades on our confusing the way a computer might simulate what a mind does and the way it might duplicate the complete causal processes whereby a mind understands, down to the last essential detail. All of that could be understood as following a program. In the thought experiment where you play the part of a computer, you haven’t been following the program that would constitute thinking in LOT; you’ve only been following a program that simulates it by delivering equivalent results. Notice that, while the computer emulates the results of intelligent processing, you emulated the computer's processing, not the intelligent processing. What if the computer had to emulate not just your appropriate purposive action but the processing by which you come to act appropriately and purposively? (A dualist should say this is impossible, but must anyone who says this is impossible be a dualist? What’s someone who says there’s a causality so novel that it directly causes meaning and understanding?)

One may intuit that a computer could act as if it had understanding, but actually not have any. But unless you're assured that the computer operates
causally equivalently to the way human beings do and still doesn't under-
stand, you cannot conclude that the connection, which the naturalizer claims
to be between action and understanding, is not as claimed. When the thought
experiment is reworked so that the computer operates causally equivalently
to the way human beings do, we fail to intuit a lack of understanding. Now
the interesting work begins; backing off from the equivalence, how much of
the causal detail is inessential to understanding, so that a computer might
leave it out and still understand?

In defense of the argument, you'll be told that the naturalizer can appeal
only to the size and complexity of the mind's program for thinking in LOT
and changes in such physical traits are insufficient to invalidate the results
of the thought experiment. If you did not understand the Chinese inputs or
outputs by mimicking the computer, you won't understand them by following
another program that gets the same result by a longer more circuitous route.
Computing is syntactic; understanding is semantic. The one does not suffice
for the other. Therefore the intentionality of the mind is not accounted for
by its being a computer. So say the critics of this form of proximate
naturalization. The naturalizer's reply is to note that the size and complexity
of the mind's program just might make the difference. As the saying goes,
"God is in the details." But if so, it's premature for us to try to settle this dispute.

It is premature. Here's a thought experiment that fits the naturalizer's
specifications for a thought experiment that might refute the naturalistic
reduction: Suppose scientists make a syntactic processor out of silicon
chips, which duplicates almost down to the last neural detail your brain's
causal pathways. Suppose they tap into your memories and load into the
syntactic processor a complete stock of your memories. Suppose they also
arrange for a way to load the processor's memories, should it acquire any,
back into your brain for you to remember. The scientists hook the processor
into your nervous system as a backup to your brain. They also installed a
switch that allows them to take your brain out of the causal loop and
substitute your syntactic processor. Now comes the experiment: You read
a book. The scientists don't tell you this, but you read all the odd numbered
chapters with your brain and all the even numbered chapters with your
syntactic processor. At the end of each chapter the scientists transfer the
accumulated memories of the chapter from the organ you used to the organ
you will use for reading the next chapter. When you finish the book, they
question you twice about your understanding of the chapters, once with your
brain in the loop and once with the silicon syntactic processor in the loop.
The scientists themselves find no difference in your understanding of the
chapters in either sequence of questionings. Well, that lack of difference to
spectators was granted by the critics of the naturalizers. Now the scientists
ask you if you feel there's any difference in your understanding of the chapters. What's the answer?

Here are two possibilities: If when your brain is in the loop, you say that there's a curious feeling about every even numbered chapter, that in retrospect it felt a bit like having read Jabberwocky, then the critics win. Alternatively, suppose you say that nothing seems different to your introspections, and you say this even with your brain in the loop. Waiting out in the hallway are all the proximate naturalizers. When they hear what you say, they pop the champagne and abandon themselves to wild celebration.

Now here's the real thought experiment: Do you here and now know which of the two possible outcomes would occur? I don't. The Chinese room argument focuses one of the issues well, but it settles nothing.

This concludes our brief survey of proximate naturalization. Negative feedback systems and computation are the key ideas we explored and defended. Together they give the naturalizer a way to reduce purpose and rationality to physical properties of physical things. What of consciousness? I beg off that one.

§4. Ultimate naturalization: there's more than one way to go about it.

Let's turn now to the other form of naturalizing, ultimate explanation. I assume that everyone will assign Daniel Dennett's philosophizing about the intentionality of the mind in an evolutionary way, in his *Darwin's Dangerous Idea*, the status of Plan A for ultimate naturalizing. By calling it Plan A I mean it's the strategy which is generally considered the most promising of success. I'm in favor of philosophizing about the mind in an evolutionary way, but I disagree with Plan A. In this subsection I want to break the hold of certain ideas that might prevent you from seeing alternative ways of carrying out an ultimate naturalization of the mind.

Dennett's Plan A can be interpreted as an inversion of the ontology accepted in Europe in the time just before Darwin, graphically presented by the pre-Darwinian cosmic pyramid (p. 64):

```
God
Mind
Design
Order
Chaos
Nothing
```

The pre-Darwinians interpreted reality as flowing from the top of the pyramid downwards. They put a level of design between the levels of mind and order. Minds create the things that display design, and order is the...
Part IV, Section Y

result of design. Dennett’s darwinizing of the pyramid accepts the level of design; he just reverses the direction of explanation. In place of the pre-Darwinian top-to-bottom order of the higher explaining the lower, the post-Darwinian posits a bottom-up direction of explanation. Top-down construction Dennett calls skyhook strategy, by which he wants us to think it’s a phony strategy since a hook cannot be suspended from the sky. Bottom-up construction uses conceptual cranes, and is not phony, since we build skyscrapers with cranes. When the level of order is understood to include natural selection from a variety of heritable traits, it has as its outcome the level of design, and when the level of design is understood to include use-functions, it has as its outcome the level of mind. The bottom-up strategy is feasible, since there are conceptual analogs of cranes. I use cranes too, but not Dennett’s, which fall down.

I’ll have nothing to say about the top or the two bottom levels, just the middle three. What status should post-Darwinians accord the level of design? Is it objectively real, or is it in the mind of the beholder merely? Pre-Darwinians opt for the former, and so does Dennett. Dennett thinks the intermediate level of design helps us understand natural selection, how the succession of its products traces a path through what he calls design space, a logical space of possible designs. The design level also represents the functionality, i.e., the usefulness or adaptation to their environments, of the things that natural selection produces. (I’ll speak of “use-functions” to distinguish them from the mathematical functions we’ve spoken of already.) So according to the design-level ontology, the heart is not just a pump-like thing, it is a pump. The ontology backs up the epistemology, namely, the claim that the conceptual framework of design yields scientifically adequate explanations of the phenomena.

Understanding the heart as a pump and consequently its evolution by selection for pumping-ability is a good instance of adaptationist thinking in biology. Adaptationism is a strategy for thinking about evolved things as adaptations to their environments, that is, their having evolved by natural selection for traits that help, and against traits that hinder, survival and reproductive success. Biologists don’t insist that the strategy pays off for all the traits of interest, just most of them. One might be an adaptationist and deny that certain common traits (for example, senescence) evolved for adaptive reasons. Other traits evolve as maladaptive side-effects of adaptations, as in the case of sickle cell anemia. Still other traits survive long after their contribution to adaptation has disappeared. For example, all vertebrate nervous systems have two pathways, one for the right side of their bodies and one for the left. The paths cross over each other just below the brain, switching sides. It’s extremely unlikely that the switch is an
adaptation for any extant species to its contemporaneous environment.

Despite these warning-signs, Dennett grounds adaptationism in metaphysics; he accords the design level the status of objective reality and thereby justifies not merely the utility of the adaptationist research methodology for thinking up evolutionary scenarios. It’s more than that; its explanations are realistic, for they reveal an objective feature of things: design. Dennett ontologizes the design level, if I understand his writings since 1991, which back away from his earlier instrumentalism. Perhaps Dennett would only think of the epistemological dimension, the giving of realistic explanations, as divided into levels, while the ontological dimension, the backing for those explanations, is more a continuum of types of being. That’s still an objectionable ontology to me. Another proponent, less equivocal in his support of the reality of the design level, is Fred Dretske.

My chief objection to the design level is that it authorizes explanations of a thing’s existence and structure by reference to its ends. It’s contrary to the naturalistic point of view, however, to countenance such teleology except in the case where the reference to an end can be reduced to an antecedent efficient cause which has that end in view, i.e., a cause which is a contentful state. We’ve seen the maximum naturalistically allowable expansion of such causes in our analysis of negative feedback systems. Dennett’s level of design is not so restricted. It’s just old fashioned teleology, no matter whether it derives from levels above it or below. The alternative I wish to develop is that the design level and use-functions, insofar as they’re supposed to exist prior to intentionality, are simply pretense and the adaptationist stance is also. What scientists see when they adopt this stance are fictions or else intention-driven systems. The case of fiction does not make the biologists who adopt it any the less productive, although they’re mentally prepared better for its frequent misguidance, if they’re aware of the fictionality. The value of this alternative of pretense, mere as-if-ness, for philosophy is in the ridding ourselves of use-functions, and a fortiori the ridding ourselves of bad explanations of the existence of mental representations in terms of them.

Let’s rethink the value of adaptationism in terms of the pretense alternative. Concerning intentionality, will adaptationism explain it? It might lead us astray. Recall how connectionist systems can create memories of nonexistent persons (section R). Representing the nonexistent and the false is a distinctive mark of intentionality—Brentano’s thesis (section G§7). I attributed this ability or rather susceptibility to a flaw in the design of the system. In other words, I think the adaptationist strategy might fail to account for some of the distinctive features of mentality. Other features that are probably byproducts of an adaptation, but not themselves adaptations,
are the egocentricity of beliefs and now-based perceptions of time, which were the focus of sections M and N. I recommend adaptationism as strongly as Dennett and Dretske do for its ability to help biologists hypothesize ultimate explanations, although I think we must be prepared to attribute some features of mentality to design flaws. Perhaps I could convince Dennett of that, but we differ fundamentally about the metaphysics he imports into adaptationism. I don't accept the metaphysical accretion to it, this pre-intentional level of design, which is Dennett's particular contribution. Virtually everything which natural selection has shaped has design flaws, and that fact shows adaptation is not real design. For example, the vertebrate eye, the traditional example of design, is flawed by having a blind spot close to the center of vision. This is no adaptation, but an instance of “the cumulative historical burden” we carry as products of eons of natural selection, as one biologist put it.\textsuperscript{15} It's an utterly mindless process and its products always betray the fact by some flaw or other.

From the design level, Dennett builds the intentional level, the mind. His chief crane is the fact of use-functions. He thinks they exist objectively at the design level. Dretske agrees. With use-functions they build representations, which are the core components of the mind level. For representations are supposed to bring with them the distinctive marks of mentality such as semantic meaningfulness. Representations are constituted by use-functions, they say. X represents Y for Z, if and only if Z uses X in a Y-representing way to achieve its goals (roughly, and the definitional circle is mine). That thesis goes beyond conceptual role semantics. Conceptual role semantics is consistent with some representations existing prior to having a role, by virtue of their carrying information into the mind. They take on the role they have in the mind because of their being found representations—something I called indicator semantics in section S§4. To say all representations are representations solely by virtue of their use-functions amounts to denying the existence of found representations. For before these representations are found they represent without being used.

Natural use-functions, if they're to be distinguished from mind-constituted use-functions, involve history in the very concept of them, in particular, a history of being selected for. (The use-functions of artifacts, in contrast, are not conceptually bound to histories.) So Dennett's Plan A posits this order of definition:

\begin{equation*}
\text{selection history} \rightarrow \text{use-function} \rightarrow \text{representation and other essentials of mind.}
\end{equation*}

Given the use-function of a sign, he can say what it's a sign of. And he can say what it's a sign of only if given its use-function. Since representations are at the level of mind and use-functions are at the lower level of design,
Dennett's construction of the aboutness of a representation is an example of a crane, building the higher from the lower.

§5. Criticism of the argument for design's coming before mind.

I wish to convince you that Dennett’s crane fails and we don’t need it anyway. Let's examine one of Dennett's defenses of his position against criticism, and see the defense fail. Dennett sees Evans's version of the twin earth experiment (see section S§3) as a challenge to his account of representations in terms of use-functions. So he reanalyzes it, but in the course of doing so he sets up a dilemma: Either buy into his position or accept a sky-hook explanation. We must follow his analysis to see how there's a third way out:

Twin earth is enough like earth to fool the ordinary person. Suppose Hume or any earthling of his generation were transported to twin earth while asleep. He wakes up on twin earth, quite ignorant of the switch in his habitat. On twin earth there's a stuff with all the common appearances of water, but it's not H$_2$O, but rather another potable, life-sustaining substance, XYZ. Hume, upon waking up, calls it “water.” He's wrong. All twin earthers call it “water”; they are right. Hume and they use the same sounds to mean different things. So far Dennett sides with Putnam. Suppose Hume stays there for the rest of his life. Dennett extends the scenario across the long time that Hume stays there. The extended scenario leads to a divergence of intuitions. There are two viewpoints to separate. There's Hume's, who's always oblivious to the switch. Yet he survives. Then there's our omniscient viewpoint, as we observe Hume. According to Dennett, from our point of view, as Hume continues his life none the wiser, the use-function of his word “water” changes from what it was on earth, so that it eventually begins to mean XYZ stuff, even though nothing at all relevant to the semantics of “water” changes in Hume himself. According to Dennett that intuition supports his use-function account of the meaning of “water.”

Some of Dennett's critics believe in “original intentionality,” that is, intentionality not built up from a level below mind. According to their intuitions, if nothing semantical happens in Hume while he accumulates experience on twin earth—for example, he does not himself decide to change the meaning of the word—then nothing happens at all to change the meaning of his word. He goes to his death on twin earth misapplying his word “water.” For the “original intentionality” folks, his word means the stuff pointed out to him on earth when he learned the word and whatever else shares the same essence as that stuff, unknown though that essence be to Hume. That the word works for Hume on twin earth is irrelevant to what
it means for him. The switch in use-function, as seen from our omniscient point of view, is irrelevant. That crane doesn't work. Hume himself must change the meaning of the word, if it's to change. So say the original intentionality folks.

Believers in original intentionality not only reject the definition of a word’s meaning and reference in terms of its use-function, they reject all ontologizing of a pre-intentional design level and its use-functions. The heart is not a pump, says Searle, and, whatever the functions of teddy bears be, real bears don't have functions, says Fodor. They've no intention-independent design or use-function. This doctrine of no design independent of mind, if true, deprives Dennett of his chief crane. I'm inclined to think the “no mind, no design” doctrine true.

I'm siding with the underived or “original intentionality” folks in my assessment of the debate over use-functions, but not with their assessment of the outcome of the twin earth thought experiment. I accept an account of it that allows Dennett the change in meaning of Hume's word “water,” but not his explanation of the change by a change in the word's use-function. To set up my alternative account of the change, I quote Gareth Evans's “The Causal Theory of Names”:

Suppose I get to know a man slightly. Suppose then a suitably primed identical twin takes over his position, and I get to know him fairly well, not noticing the switch. Immediately after the switch my [mental] dossier [on the guy] will still be dominantly of the original man, and I falsely believe, as I would acknowledge if it was pointed out [to me], that he is in the room. Then I would pass through a period in which neither was dominant; I had not misidentified one as the other, an asymmetric relation, but rather confused them. Finally the twin could take over the dominant position; I would not have false beliefs about who is in the room, but false beliefs about, e.g., when I first met the man in the room.

To understand this quotation, think of our beliefs as connected both physically and semantically to their subject matter, physically by the subject matter somehow causing the belief to come into existence and semantically by the belief being about the subject matter. On the semantic side, beliefs and other attitudes about people are organized into mental dossiers about each of them. All the beliefs in a dossier at a time are at that time about the same person. That person's name is the label on the dossier. On the physical side, occasional confusions cause misfilings, but the dossier is dominantly caused by one person, and so all the beliefs in the dossier are about whichever individual is the cause of most of the beliefs in it. Now consider the dossier as it exists over the course of time: The idea of a single dossier, collecting filings as time goes by, changing from
containing beliefs dominantly caused by one guy to containing beliefs
dominantly caused by another guy. Consequently, all the beliefs in the
dossier change semantically, from being about the first guy to being about
the second guy. Yet it remains the same dossier, so that the label on the
dossier changes its meaning. Well, that's the intuition pump that makes me
agree that Hume's meaning changes, as Dennett thinks.

Use-functions have nothing to do with this story, however. To the junk
yard with Dennett's two-bitsers in Panama, and all the other paraphernalia
he trots out to show that changes in use-functions are at the bottom of this
change! Having mental dossiers on people, water, and horses makes sense
to me, and they are all we need to accommodate the intuition of meaning
change. I reject the intuitions of the original intentionality folks on this
thought experiment, and I don't agree with Dennett that it's use-function that
matters. It's simply a matter of belief-count, based on the beliefs' causes.13
When the count of beliefs caused by twin water in Hume's water dossier
exceeds the count of beliefs caused by water, the meaning of Hume's word
“water” changes to mean twin water.

§6. Plan B for implementing the naturalistic project.

Dennett offered us the philosophical Darwinian's “Plan A.” I suggest we
consider a “Plan B” for ultimate naturalization of the mind. The chief
feature of Plan B is to naturalize the mind by appeal to evolution without
ontologizing the design level and without accepting as true the appearances
of natural functionality. Objectively, there are no mind-independent designs
or use-functions. That admission entails that we cannot use functionality as
a crane for introducing representations. It would be circular, deriving mind
from the mental.

Who needs a crane for introducing representations? Nobody. Consider
the bimetal in a thermostat of a system for regulating temperatures by way
of negative feedback loops. Do changes in the shape of the bimetal repres-
ent changes in temperature only because of the bimetal's function in the
system? Of course not. Changes in the one represent changes in the other
regardless of the use-function. In fact the very functionality of negative
feedback systems depends on the antecedent existence of representations.
The temperature regulator wouldn't work, if a bimetal's states did not
already represent temperatures. Now extrapolate to all the naturally occur-
rering systems. According to Plan B, representations are already in the pre-
life world waiting to be found and exploited by life to serve the fitnesses
of organisms. If bimetals had existed in nature, natural selection would
most likely have exploited them eventually for temperature control, as it has
exploited magnetite to help some species of migratory birds to navigate the globe. Magnetite free to move around represents north-south directionality, whether or not anyone uses it for a compass. Plan A had the story reversed. The order of explanation is:

representations —> selection for exploiters of representations, in tandem with the use-functions which they endow things with.

More fundamentally, if there weren't representations of states of the environment and of goals, nothing would be good for anything, and there'd be no use-functions.

Plan B for the ultimate naturalizing of the mind is to see how the mind level can be built up using cranes from the levels of mere life and indeed from the pre-life level, directly and without benefit of a mediating level of design and functionality. One ingredient of Plan B is the fact that life does not have to invent representations. It simply has to find them and exploit them. What life does invent is purposes, which are representations of organisms' goal states. Biological theory posits the value called fitness, in terms of which we appraise organisms' purposes. Biological value, namely fitness, and the purpose to actualize it are there on the ground floor of life, mere life, along with representations. They're what's available for the construction of minds.

Natural selection, according to Plan A, entered into the definition of everything mental, because it entered into the definition of use-function. According to Plan B, natural selection does not enter into the definition of use-function or indeed of anything mental at all, either directly or indirectly. Natural selection only makes minds; what minds are is a distinct matter. I defined mind in the Introduction without appeal to natural selection. The adaptationist stance, with its method of reverse engineering, is a frame of mind that helps us understand how natural selection makes minds, but it doesn't reveal a level of reality consisting of pre-mental functions. Instead it reveals that all the ingredients of mentality are there from the start of life. Think of life not as pre-mental but proto-mental.

You may feel some residual doubts about the viability of Plan B. Let me try to allay them. First, there's something odd in saying there are representations in nature. It amounts to allowing two ways of describing material things. A representation is first just something physical with measurable intrinsic properties. Secondly, a representation can be described as containing propositional content, namely, the information it contains about something else. For example, a particular coiling of a bimetal makes it represent that the temperature of the ambient air is 300 degrees Kelvin. Nevertheless, I see nothing wrong in having this mode of description of material things available, even though science itself makes most headway
using the language of measured intrinsic properties.

One thing that gets in the way of thinking of the mind naturalistically is the suspicion that a naturalistic account of misrepresentation is very difficult, if not impossible. A proponent of Plan A might grant me my found representations, but insist there's a gulf between the found representations that carry information naturalistically and the made representations that have semantic meaning. The gulf is so wide that falsehood is impossible on the naturalistic side. Information must be true or else it's not information; who can deny that? Proponents of Plan B are then without the conceptual resources to account for falsity, for the only way to bring it into the picture is to establish a concept of semantic meaning for made representations, and that requires the prior notion of use-functions.

This objection, popularized by Dretske, is widely considered to be fatal to Plan B, but is it not based on a false dichotomy? Consider the connectionist memories in a Hopfield net, as described in Section R. The net connects feature detectors. The exogenous activation of some of them leads to the activation of the whole net, often in such a way that the activated set represents the presence of people truly, but sometimes in a way that represents no actual person. It's a representation nevertheless, for in either case the result can rise to consciousness as a memory-thought or utterance, just because people cannot exploit the true ones for a certain purpose without their risking getting false ones through no fault of interpretation. Granted there's a processing that leads to a continuum of enhanced meaning; nevertheless it seems simply stubbornness always to insist on a later stage of processing as the onset of meaningfulness, and that prior to that stage use-functions have already come into being. The state of the net has a perfectly good meaning in terms of the same feature detectors that are activated in both cases by the components’ interactions, independently of the fact that the organism processes the state further and acts successfully in the former case and unsuccessfully in the latter. I hope your intuitions agree with mine that the falsity is there in the representation independently of there being any use of it. If Dretske were to shift his dichotomy so that all examples of this sort were on the side of the gulf where there's no semantic meaning and so they were only mimicking or presaging genuine false representations and representations of the nonexistent, one would have to wonder if the debate had become merely verbal.

When it comes to explaining how Hume's word “water” changed its meaning gradually as he lived out his life on twin earth, I incorporate Evans's dossiers into my Plan B. Proper names and names of natural kinds are good candidates to be labels on mental dossiers, since there should be a division of labor between the label and the information. Hume's dossier
on water came to be predominantly filled with material based on his acquaintance with water's twin. So that's what the dossier's label came to mean.

Another thing that gets in the way of thinking of the mind naturalistically is that we tend to see all representation on the model of communication. According to Plan B, however, receivers of information just have to find it in the world, and there do not have to be intentional senders of the information. Plan B has the order right. Getting it comes before dishing it.

A final worry: In the course of evolution, the primate brain latched onto this representational way of describing its own states, and perceiving them by introspection. The primate brain got it right to that extent. Brains exploit representations in nature because organisms are structured to carry the propagation of the information into their interiors so that internal states are representations too. There they get tied into negative feedback mechanisms. Selection hones the interior spread of representational states for fitness, and so for verisimilitude and appropriateness.

One might object that the brain has trouble seeing its representational states as being states also describable in the intrinsic terms that are measurable and materialistic. That's because, having found one way of dealing successfully with its own states, there was no evolutionary pressure to find still another way. The difficulty translates into an argument from ignorance: “I do not know a second way; so there is no second way to be known.” That's something the primate brain got wrong.

Summarizing Plan B:

- There's no ontological level of design. Natural selection shapes only close imitations of design, but always with some screwball defect that reveals the burden of history.
- Intentionality itself shows many instances of these nonadaptive screwball defects. Indeed some of them are the chief identifying characteristics of intentionality, like the ability to name nonexistent objects.
- The chief crane for constructing the mind is the prebiological fact that representations exist in nature. Life simply exploits that fact to produce minds.
- The representations found in nature independently of minds exhibit all the marks of intentionality, so that the traits that distinguish mind from non-mind are less fundamental traits, and mostly matters of degree.
- Consequently mind is not so fundamental an ontological category, far less fundamental than life is.
- Use-functions are features of organisms emergent from life's
exploitation of the representations it finds in nature. The representations make use-functions possible, not the other way round. There's one more feature of Plan B, which I'll elaborate on in our last section:
• The techniques of philosophical analysis can reveal the evolution of intentionality.
If this is so, then, since analysis is inherently a conservative procedure, perhaps ultimate naturalizing fits folk psychology very well.

Notes

3. Gilbert Harman argues that intentions-to-act are self-referential in the sense that, for one to intend to act, one must see that one's intention will cause the act intended and one must intend that causation. If that's so, I've not given a complete analysis. See part 1 of his “Desired Desires” in R. G. Frey and Christopher W. Morris, eds., *Value, Welfare, and Morality* (1993) 138-157.
5. Contemporary theoretical physicists speculate that the information that's really out there, and which physics must recognize, is conserved. How it would relate to ordinary information is unclear to me. What is clear is that information ordinarily so called is not conserved.
6. The figure is copied from my two essays, “Gaia = Maya” and “Essay on Nature’s Semeiosis.” It was adapted from my “Purpose, Feedback, and Evolution” *Philosophy of Science* 48 (1981) 200, where it was adapted from Herbert Simon, “Applications of Servomechanism Theory to Production Control” in his *Models of Man* (1957) 219-240.
7. The mathematical treatment is derived from Jay Forrester, *Principles of Systems* (1968), ch. 2.2 on first-order systems.
11. See John Searle's article on himself in *Blackwell Companion to the Philosophy of Mind*. 


14. I may be misinterpreting Dennett as being a scientific realist about natural design, and he may only be an instrumentalist. His articles in the 1970s were instrumentalist. It seems to me, however, that since the late 1980s he has tried to find a middle position between realist and instrumentalist theories of theories. See his “Real Patterns,” *Journal of Philosophy*, 87 (1991) 27-51. To evaluate the coherence of this attempt would take us into the philosophy of science.


18. Well, perhaps not “simply.” Evans calls sheer count “crude” and proposes weightings of entries on various dimensions (p. 221).