Making a Thinking Robot

Bill Grundy
Senior Honors Thesis
Symbolic Systems Program
Stanford University
Professor Fred Dretske

June 5, 1991

Introduction

Yesterday, as I was playing 3-D tic-tac-toe on my personal computer, my roommate described to me the computer's strategy: "He tries to get three corners on the top or bottom level. If he does that he guarantees a win." Undoubtedly, my roommate intended his description to be metaphorical, a sort of computational anthropomorphism. Few people would seriously entertain the possibility that my PC has the capacity to maintain beliefs or desires. Most people intuitively feel that the computer cannot truly want to win the tic-tac-toe game in the same way that its human opponent wants to win. Nor can the computer believe that a certain strategy will guarantee a win.

However, intuition is not always a trustworthy guide. Even the outward appearance of a computer can strongly affect our judgement about its internal functioning. Take away the computer's metal casing and sharp edges. Enclose it in a soft exterior and give it a coat of fur and an adorable, expressive face. Then try to dismantle it. The task may prove difficult if the fuzzy computer shrieks when you cut its skin, if it scuttles away on motorized wheels, if it bleeds simulated blood. Much of the popular culture fear of "thinking computers" undoubtedly stems from a distaste for the computer's unfriendly exterior, rather than from a firm understanding of why computers can or cannot think.

Flakey the Robot

This, then, is the question this paper will address: Is it possible to create an artifact that has beliefs and desires similar in kind to our own? Although the argumentation will be primarily philosophical in nature, the paper stems from practical experience. In the summer of 1990 I interned at SRI International where I helped program Flakey the Robot. Flakey is a black, wheeled octagon about three-and-a-half feet tall. It is powered by two golf-cart batteries and carries a powerful computer on board as well as a packet radio for communication with off-board computers. Flakey "sees" by means of a ring of twelve Polaroid sonars around its base, a laser light stripe, and a video camera.

The last day of my internship coincided with SRI Family Day, so I had a chance to test in a dynamic environment many of the behavior routines I'd written. As Flakey wandered around a room crowded with rambunctious children and worried parents, I'm sure that many of the observers were convinced that something profound was occurring within the robot. When Flakey spoke, "I am six years old. How old are you?" the children would reply truthfully and try to continue the conversation.

But I found myself entirely unimpressed. Understanding most of Flakey's internal functioning made the final product seem less like hocus-pocus and more like simple mechanics. Flakey's "artificial intelligence" existed only for those who didn't understand the internal workings of the robot. The experience forced me to reconsider my conception of human intelligence. For me, Flakey's relatively complex behavior did not appear to be cognitive because I understood it too well. Perhaps, if psychology and neurology can be combined, a similar reduction may be made for human intelligence. Or perhaps Flakey lacks some essential property — a certain physical configuration or a type of history — which prevents the robot from having beliefs and desires. This paper is an attempt to find that property, to determine what would be required to make Flakey think.

Definition of "Intentionality"

The first step in the search is to label what we're looking for, if only for the sake of discussion. The term "intentionality" describes one important facet of cognition — the property of being about something else, of having "a direction upon an object" (Brentano qtd. by Chisholm 201).[1] This property of directedness is characteristic of beliefs, desires, and many other cognitive predicates: a belief can't be a belief unless it has a content, that is, unless it is a belief about something. Similarly, a desire is always a desire for something.

Making a system intentional is necessary but may not be sufficient to make the system cognitive. In order for a system to think, it must be able to think about things. However, many mental phenomena — such as pain and pleasure — are not intentional at all. Pain simply is; it has no intrinsic direction. Pain may cause you to pull your hand away from a flame, but the movement results from a belief about the cause of pain and a desire to avoid further pain (Unless the movement is a reflex, in which case it doesn't involve cognition at all). A system that experiences only non-intentional states clearly would not be a thinking system. Therefore, the search for necessary conditions for mentality can begin with a search for intentionality.

To a large extent, the way that search is conducted depends upon the searcher's theory of mind. There exist several distinct characterizations of intentionality, each with its own philosophical underpinnings. In order to determine how to build intentionality into Flakey the Robot, we have to begin by figuring out what sort of a thing intentionality is. This entails adopting a coherent, workable metaphysics.

Dualism

The most straightforward characterization of intentionality aims at explaining our use of verbs such as "believe," "desire" and "intend" by simply adopting intentional language in order to describe intentionality. This approach is essentially dualist, since it does not allow for the descriptions of mental states in non-mental language. The dualist attempts to create a complete theory of mind using only intentional terms. This is the commonsense method, since we typically explain intentional behavior in intentional terms: "Jeremy wanted to stay at home because he believed that it would soon rain."

The dualist approach makes the study of intentionality very difficult, if not impossible. Since, according to this theory, intentionality is completely non-physical, one can never determine with certainty Flakey's beliefs, or even whether Flakey has beliefs. In fact, even if Flakey were a human being, the only way to find out whether Flakey believes that, for instance, there is a wall four feet in front of it would be to ask Flakey. And even if Flakey says "Yes, I believe that there is a wall four feet in front of me," an outside observer would still not know for certain whether the report corresponds to an actual belief. This problem of Other Minds arises from our own inability to gain access to any mentality other than our own. For the dualist, simply programming the robot to say "Yes" at the right time is clearly insufficient to bring about belief. Believing requires more than simply saying "I believe": it requires that the believer actually have a

[1] "The term intentionality is a technical term drawn from Medieval philosophy . . . Although there is a relation between this term and the term *intentional* that is a derivative of intend, the two should not be confused" (Bechtel 40).

belief. And since beliefs are, for the dualist, non-physical entities, the study of intentionality cannot even begin.

Of course, just because dualism makes studying the mind difficult does not mean that dualism is incorrect; however, it does imply that a scientific study of intentionality must begin elsewhere. The metaphysics assumed by much of modern science is materialist; therefore, some form of materialism seems appropriate for the study of Flakey's intentionality.

Behaviorism

One type of materialist, the behaviorist, claims that intentional vocabulary is simply a means of describing a system's behavior. According to the behaviorists, intentional phenomena are external, behavioral events, rather than internal ones. Therefore, a true, scientific account of behavior would discard all talk of beliefs and desires in favor a stimulus-response vocabulary. For the behaviorist, if Flakey acts like a robot with beliefs and desires, then Flakey is a robot with beliefs and desires.

Unfortunately, the behaviorist agenda has proven to be pragmatically untenable. Although such experimentation has yielded invaluable data, "these gains have been achieved independently of — and, in many instances, in spite of — the theories the experiments were intended to confirm or disconfirm" (Dennett 1969 33). Problems arise when attempting to describe, for instance, experimental results involving learning. A child learning to work arithmetic problems does not learn a sequence of physiological responses; the actual physical response may differ each time the child solves a problem. However, the mathematical content being manipulated is similar. Because behaviorist explanations do not allow for thoughts, beliefs and inferences, such explanations provide no means by which the content of learning may be characterized.

Some forms of behaviorism are less dogmatic, but still open to the same criticism. Methodological behaviorism, for instance, allows for the existence of beliefs and desires but still excludes intentional vocabulary from scientific discussion. The methodological behaviorists claim that intentional phenomena either do not cause behavior or at least are unnecessary for a scientific explanation of behavior. They thereby avoid making a definite metaphysical claim while maintaining their ability to conduct scientific research. Unfortunately, their approach leaves unresolved the descriptive problems faced by the eliminative materialist. Even if the methodological behaviorist allows that beliefs exist, if the behaviorist refuses to include those beliefs in descriptions of learning behavior, then the power of the explanation will be severely limited.

Intentional Stance

Closely related to methodological behaviorism is the adoption of the intentional stance. This approach also makes no definite claim regarding the existence or inexistence of intentionality. Adopting the intentional stance simply means admitting that the intentional vocabulary of beliefs, desires, intentions, etc. is a useful tool for describing the behavior of complex systems. Although one may choose to describe nearly any system intentionally, such descriptions only seem appropriate for certain types of intuitively intentional systems. Thus one may describe a water faucet intentionally: when I turn the handle, the faucet believes that I want water to be dispensed, knows how to dispense water and wants to fulfill my wishes. Such a description, though clearly implausible when applied to faucets, becomes essential in describing more complex systems such as human beings. That was, after all, the primary lesson learned from eliminative materialism: explaining behavior requires at least some intentional vocabulary. Adopting the intentional stance is an admission of that lesson.

Many computational systems are best described using intentional language. For instance, the tic-tac-toe strategy of my personal computer lends itself to a description in terms of the PC's desire to win and its beliefs about strategies. Even more so, Flakey seems amenable to intentional description: for one thing, the robot appears to have a genuine aversion to walls. Certainly to the children at SRI Family Day, the only plausible description of Flakey's behavior would have to be framed in terms of the robot's beliefs and desires.

However, an intentional description of Flakey does not imply that Flakey is intentional. Unlike dualism and eliminative materialism, adopting the intentional stance does not entail a strong psychological thesis. The intentional stance concerns itself with language rather than behavior. It does not specify whether intentional language describes a physical phenomenon, an epiphenomenon or nothing at all. Perhaps intentional language simplifies the true but more complex physical descriptions toward which science aims. Or perhaps intentional language picks out some essential psychological function. Adoption of the intentional stance leaves the possibilities open. Thus, while maintaining the intentional stance toward Flakey, one may either agree with the children in saying that Flakey actually has beliefs, or one may agree with the programmer in saying that Flakey only superficially acts like a robot with beliefs. Neither approach contains a contradiction.

The intentional stance approach accurately describes the way we use intentional language, but it provides little or no guidance in the search for intentionality itself. People use intentional language to describe intentional systems such as human beings, as well as to describe non-intentional systems such as tic-tac-toe-playing computers. People even alternate in their descriptions of a given system: the computer acts intentionally while it plays tic-tac-toe but becomes non-intentional when it halts due to a system error. We say the computer's beliefs cause it to carry out the correct game strategy, but those beliefs do not cause the computer to crash when it opens too many files at once. If intentionality is to be more than simply a mode of description, then the language of beliefs and desires cannot serve as a criterion for intentionality.

Representation

The search for intentionality thus appears to have reached an impasse. Since what we describe as intentional behavior may accompany both intentional and non-intentional systems, we need to find some characteristic which occurs in all and only the truly intentional systems. This defining characteristic may be taken as representation. The most important characteristic of beliefs and desires, at least from the point of view of the believer or desirer, is that they represent. "Whatever else a belief is, it is a kind of thing of which semantic evaluation is appropriate" (Fodor 1). Not only is such evaluation appropriate; it is necessary. A belief is not a belief unless it is believed. This is what seems to have been missing in the search for intentionality — the other half of the intentional relation. We're not simply looking for aboutness; we're looking for aboutness for some agent.

Resemblance

Just as the term "intentionality" inevitably runs into confusion with the verb "intend" and with the linguistic term "intension," the word "representation" has several everyday meanings which should be excluded from this discussion. First, representation is not resemblance. A painting of Kermit the Frog is a representation of Kermit the Frog, but it does not represent Kermit because it resembles Kermit. After all, a painting may represent something without resembling it. If a four-year-old child draws a purple and orange splotch on a sheet of paper and then announces that she has drawn Kermit, then, at least for the child, that splotch represents Kermit the Frog. This type of assigned representation is the way words represent. For instance, when Jim Henson first pasted together his green froggy-looking Muppet and said, "I think I'll call you 'Kermit the Frog,'" he assigned a representational meaning to those three words — "Kermit the Frog" — in the same way that the four-year-old child assigned meaning to her purple and orange splotch. Nothing in the words themselves resembles Kermit. Similarly for beliefs: I believe that Kermit the Frog is green, but that does not imply that somewhere in my head (or wherever beliefs exist) there is a belief that looks like a green frog. If beliefs did resemble their contents, then it would be impossible to believe, for instance, that honesty is a virtue, since there is no way for a belief to resemble either honesty or virtue. Representation does not mean resemblance.

Assigned Representation

Furthermore, representations in the human mind differ from the painting of Kermit the Frog in the way they gain content. If all that were required in order to represent was some kind of dubbing event (like "I think I'll call you 'Kermit the Frog"'), then Flakey would have little difficulty in becoming representational. In fact, even my PC would represent, since my roommate has already specified the representational content of its beliefs about tic-tac-toe. The kind of representation that occurs in humans must be of a different type.

The Chinese Room (John Searle)

Most significantly, human-style beliefs and desires must represent for the system of which they are a part. This is the point of John Searle's famous Chinese Room example: representation of the simple, assigned sort is insufficient to guarantee genuine thought. In order to be significant, the representation must represent for the system itself. In Searle's example, an English-speaking, human subject is locked in a room and given three sets of Chinese characters, along with rules in English for correlating the various sets of characters and producing output characters. Because the subject knows absolutely nothing about Chinese, the input and output characters appear to him as meaningless squiggles. However, unknown to the human subject,

> the people who are giving [him] all of these symbols call the first batch "a script," they call the second batch a "story," and they call the third batch "questions." Furthermore, they call the symbols [the subject] gives them back in response to the third batch "answers to the questions," and the set of rules in English that they gave [him], they call "the program." (418)

Eventually, the human subject develops such skill at manipulating these various sets of symbols that people outside the room cannot distinguish his output from that of a native Chinese speaker. Searle argues that even though the human subject appears, from the outside, to have the ability to represent, he does not actually do so since the subject has no access to the content of his representations. If the subject's beliefs have no meaning for the subject himself, then he can't really be said to hold those beliefs at all. Thus it makes no sense to say that I believe that Kermit the Frog is green but am unaware that Kermit is green. Cognitive representation is a relation which includes both a semantic content and an agent.

So far, we know that the kind of representation we're interested in is not a resemblance relation, that it does not attain as a result of some external assignment of meaning, and that it must have meaning for the representing system. This description of representation is sufficient to begin evaluating some theories of representation.

Physical Characterization of Representation (Dan Lloyd)

In *Simple Minds*, Dan Lloyd attempts to discover the basis of representation by describing a representational system of minimal complexity. The result is a definition of representation in terms of three conditions: multiple channels, convergence, and uptake. Lloyd presents these three conditions as both sufficient and necessary for representation; however, he bases his claims almost exclusively upon examples rather than arguments. Unfortunately, the examples he provides are not intuitively representational, and even if they were, Lloyd provides no explanation of why the characteristics of the simple minds he describes must be common to all representational systems.

Pre-theoretic Constraints

The first chapter of *Simple Minds* outlines seven constraints to which any representational system must conform. They are accuracy, focus, articulation, cognitive role, evolutionary explanation, asymmetry of relation, and reduction (Lloyd 12–19). The first three constraints concern picking out items in the world. These constraints do not require perfection: representational systems are not required to always be correct and perfectly detailed in what they represent. However, any representational system must be able to maintain a fairly robust correspondence between its internal representations and the objects being represented. The fourth constraint, cognitive role, is simply the requirement that the representation be causally efficacious in some relevant way. The fifth constraint of evolutionary explanation requires not that the system have an evolutionary history, but only that it might have had such a history. Finally, the last two constraints require that there exist a directionality to the representation relation ("A line drawing may represent you, but you do not, by that token, represent the drawing" (Lloyd 17).) and that the theory of representation be stated in reduced, non-representational terms.

Lloyd's Definition of "Representation"

Lloyd derives his definition of representation through a description of the evolution of imaginary vehicles. He begins with a simple, two-wheeled Braitenbergian vehicle (following Valentino Braitenberg) which he dubs "Squint." This most primitive form has a single, photo-receptive transducer which powers a small motor. Its sole behavior is the avoidance of light: the vehicle always moves forward until it reaches darkness. Lloyd describes the various problems associated with such a simple vehicle, such as its unreliability and lack of perceptive discrimination. To make Squint more reliable, Lloyd proposes the multiplication of transducers. Redundancy of input allows cross-checking and cancelling of errors. Then, Lloyd upgrades the photo-receptors so they become directional. This improvement allows two transducers to point across one area, providing the vehicle with specific information about the distance of the light source.

The final product of this pseudo-evolution exhibits Lloyd's three definitional characteristics of representation. The first is the multiple channel condition, which requires that the representation depend upon two or more afferent events. Along with the multiple channel condition comes the convergence condition, which simply stipulates that the several events coming through multiple channels must converge to the representation of one, most probable event. Finally, in order to meet his "pre-theoretical requirement that the content of a representation make some difference to the system" (Lloyd 66), Lloyd adds the uptake condition, which requires that a representation cause either "another representation or a salient behavioral event" (Lloyd 64).

Lloyd presents these three conditions as merely sufficient for representation but treats them throughout the book as if they are also necessary. The formal outline of the three conditions begins, "A natural event r is a representation if it meets these conditions:" (Lloyd 64). In other words, the multiple channel, convergence, and uptake conditions are sufficient for representation. However, Lloyd later re-states the multiple channel condition: a representation depends "upon two or more [events] and will not occur otherwise" (Lloyd 65). This implies that at least the first condition is necessary for representation. The second condition, convergence, follows logically from the first. Given that multiple channels are required for representation to occur, it would be difficult to represent anything without convergence as well. A non-converging, multiple-channelled system would only be able to represent amorphous possibilities, each corresponding to one input channel or a combination of several channels. Finally, the uptake condition must also be necessary, since it was introduced to account for a pre-theoretic requirement. The uptake condition simple restates Lloyd's intuitive requirement that representations not be causally inert (Lloyd 17–18). A system which does not fulfill the uptake condition would not meet the original description of a representational system. Thus Lloyd presents all three conditions as jointly sufficient and individually necessary for representation.

Lloyd's Definitional Characteristics are not necessary

However, Lloyd fails to support his implicit claim that the characteristics of the vehicle are necessary for representation. Certainly the method by which Lloyd derives the machine's characteristics does not imply that the multiple channel constraint is necessary. Lloyd describes the evolution of Squint as follows:

> In the pursuit of reliability, we might upgrade the basic material of Squint's circuits. Suppose, however, we are stuck with something like the raw material that was Squint's original endowment. (Lloyd 54)

Thus the redundancy embodied in Squint, which later becomes a part of the definition of representation, is contingent upon the fallibility of Squint's circuit material. We might suppose, however, that either Squint's circuits were far more fallible, in which case even the moderately redundant circuits in the final product would be insufficient for reliability, or that Squint's circuits were far more reliable, in which case the redundancy would be superfluous. In short, if the described redundancy is really just a means of overcoming error, then it certainly is not necessary. There are other ways to be reliable.

Flakey uses a more complicated transducer, a laser light-stripe. This transducer has several parts. One half consists of a laser that shines through a diverging lens to create a horizontal stripe of infrared light. The other half is a video camera whose lens is capped with a finely-tuned filter. The filter only allows wavelengths within a few nanometers of the laser's wavelength to pass through the lens. Because of the high coherence of the laser light, the resulting line, as viewed by the video, shows distinct edges. Because of the parallax between video and laser, the height of the laser line on the video image is correlated with the distance of upcoming obstacles. Thus the laser data, transmitted through a single camera lens, contains far greater and more accurate information than could ever be provided by, for instance, the battery of twelve Polaroid sonars which Flakey carries around its base. When Flakey only uses the twelve sonars, the robot gets poorly correlated, inaccurate data, but it does so through many different channels. Therefore, according to Lloyd's definition, Flakey can only represent by means of its sonars.

That seems counterintuitive. A robot using only laser light data will fare better in behavioral tests than an identical robot using only sonar data. If what Lloyd means by representation is anything like the commonsense usage, then it seems that the laser-equipped robot represents objects just as much as the sonar-equipped robot. And the former robot does so with far greater accuracy. Especially if Lloyd wants his theory of representation to be causally linked to behavior (by means of his uptake condition), then it seems that the decision as to when representation is occurring should depend upon the relationship between representer and representee rather than upon the internal physical make-up of the representer.

If reliability does not necessitate the multiple channel constraint, then perhaps that constraint might be justified by drawing an analogy with evolved neurology. Human neural activity is characteristically massively parallel. While such an analogy would not alone justify the inclusion of multiple channels in the definition of representation, it might be taken as a good indicator that parallel architectures are useful in representing.

However, biology is not the best teacher in this case, since the parallel circuitry of the brain is probably evolution's attempt to overcome the relatively slow speed of the neuron. Since a neuron's firing speed only allows for a few hundred connections to occur before a behavioral response is needed, the brain compensates by energizing hundreds of neurons at once. In a silicon-based computer, however, the computing speed can be thousands or even millions of times faster than neurons. Therefore, computers bypass the need for a multiple channel condition. Thus the existence of multiple channels in the neural architecture of evolved organisms is contingent upon the speed at which neurons fire; the parallel architecture is not an essential aspect of the representing system. Thus neither the need for reliability nor an analogy with evolved organisms can provide a justification for the multiple channel condition.

Lloyd's Definitional Characteristics are not Proven to be Sufficient

Lloyd bases the sufficiency claim upon the fulfillment of all of his metatheoretic requirements for representation. According to Lloyd, since the multi-channelled version of Squint meets these seven constraints, it is a representing system: "It doesn't do much, but it does what it needs to, by the lights of the metatheory" (Lloyd 62).

Granting that Squint does meet Lloyd's metatheoretic constraints, however, does not guarantee that the vehicle has the ability to represent. Lloyd has not guaranteed that his description exclusively picks out all and only those systems which can represent. I might send some hunters on a tiger hunt with only the description, "Tigers are six feet long, with razor-sharp teeth, and stripes." But I shouldn't be surprised when they return with a lumberjack's saw painted in black and white stripes. The hunters would be wrong in calling the saw a tiger, but they would be even more mistaken if they then expanded the definition of tigers to include the requirement that tigers have handles on both ends. This is, in a sense, what Lloyd has done.

The first question, then, is, "In what sense is the multi-channelled Braitenbergian vehicle like the striped lumberjack's saw?" The hunters on the tiger hunt didn't know about at least one essential tiger characteristic — that tigers are living creatures. Similarly, there may be an essential characteristic of representation which Lloyd has overlooked, since there is no reason to suppose that Lloyd's metatheory constitutes an exhaustive description of representation. If it is not, then simply fulfilling that description will not determine the sufficient conditions for representation. Lloyd's methodology therefore provides no guarantee that the multi-channel vehicle he describes can actually represent.

Physicalism in General

In fact, the criticism of Lloyd's physicalism can be generalized to show that any account of representation which specifies only the current physical characteristics of representational systems cannot succeed. It would be difficult to prove that a given physical trait is necessary for representation to occur. Representational creatures from Mars or beyond need not necessarily share any physical characteristics with humanity. Intelligent alien life arguably would require some physical instantiation. But the alien's form seems unlimited: there is no reason to suppose that martians have multiple channels, that their "neural" communications are electrical or even that their chemistry is carbon-based. Whatever specific intrinsic property a physicalist theory specifies as defining representational systems, it should be possible to posit a representational alien without that property.

Need for a Historical Condition

The problem, then, is this: Lloyd's simple, three-part definition of representation is unconvincing. On the one hand, it doesn't seem sufficient, since Lloyd's evolved systems do not obviously represent. On the other hand, the definition cannot be necessary since it is framed solely in terms of the current physical characteristics of his evolved system, characteristics which are contingent upon the evolutionary history of Lloyd's vehicles. What's still required, then, is a trait or set of traits that are characteristic of all and only the truly representational systems. But if that trait is not physical then the theory will fall into the dualist circularity described above.

One approach which avoids this dilemma looks at the history of the representational system instead of the current physical system itself. Such a teleological theory of representation is still materialist, since it does not posit non-physical mindstuff, but the theory avoids Lloyd's dilemma since a certain type of history does seem necessary in order for a system to represent. Certainly, meaning in human language depends not upon the actual sound uttered but upon the history of that sound. An analogous argument can be made for representation: in order for an internal structure to mean something to the system, that is, to be a belief or desire, it must have some historical connection to its own content. If the semantic character of a belief has nothing to do with the belief's being there, then it shouldn't really be called a "belief" at all, since the system supposedly doing the believing has no pragmatic association between belief and content. An alien creature would have no chance of accurately representing a Corvette unless the creature had some historical link to Corvettes. Even if the alien could think about an item exactly similar to a Corvette, it would not actually be thinking about a Corvette: it would merely be thinking about something just like a Corvette. Similarly, in order for Flakey the Robot to think about Corvettes at all, the robot would require some historical connection to Corvettes. The teleological theory of representation begins with this necessary connection between representation and content.

Papineau's Evolutionary Condition

Clearly, the teleological theory of representation must specify the historical connection more definitely than simply as a connection between content and representation. Otherwise, every object would have the ability to represent all of its causal ancestors. David Papineau, in Reality and Representation, suggests that evolution serves as the historical basis of representation. He claims that representation should be understood in terms of the biological functions of beliefs and desires. Thus since, by Papineau's account, nothing in Flakey the Robot has a biological function, the robot lacks the ability to represent anything at all.

For Papineau, representations depend upon the biological significance of their contents, which in turn depend upon the history of the representing system. In order to clarify the term "biological significance," Papineau turns to natural selection. By the selectionist account, the function of, for instance, a narwhal's tusk is not determined by the tusk's immediate capacities in helping the narwhal to survive. Indeed, the tusk doesn't seem to serve any such purpose for individual narwhals: it is merely a secondary sex characteristic, like a beard in humans (Lopez 131). Instead, the tusk's function depends upon the history of narwhals, upon the survival of those narwhals with tusks of this particular type. Perhaps a narwhal with one huge tusk is more likely to attract a narwhal of the opposite sex than its similar, tusk-less counterparts. By Papineau's account, then, the biological significance of the tusk is social. It depends upon the selective evolutionary history of the narwhal.

The selectionist account works well for an apparently useless appendage like the narwhal tusk: without an account by natural selection, the tusk seems functionless. However, Papineau extends his argument to include any part of an organism, even those parts which serve an immediate and obvious biological purpose. Thus the function of a octopus's ink does not depend upon the octopus's use of that ink to befuddle would-be attackers; it depends upon the survival of the octopus and its ancestors by using the ink as a defense mechanism. Even if the octopus were to begin using its ink to flush prey out of tight corners, Papineau would not call that behavior the ink's function until the prey-flushing capacity had given the octopus a selective advantage over its competitors. Furthermore, an octopus in an aquarium who never had occasion to squirt ink at all would still contain ink whose function would be to confuse predators. Thus the ink's function depends upon its history rather than its current contribution to the octopus's welfare. This distinction comes out clearly in one of Papineau's examples:

> Consider . . . a little green bug who coagulates by cosmic happenstance, and just happens to get a rudimentary limb suitable for putting food into its rudimentary mouth. Would one want to say that the little green limb was there in order to help the creature feed itself? I would say the creature was just lucky to have the limb. (74)

If, by an even greater cosmic happenstance, this bug were to find another little green bug with which to mate, and if the bugs' rudimentary limbs enabled the little green bugs to outlive their competitors, the medium-sized brown bugs and the very tiny blue bugs, then presumably Papineau would grant that the limbs of the little green bugs' descendants have a biological function. But without such a history of selection, considering only the limbs' current abilities, Papineau's account describes the limbs as functionless.

Evolution of Beliefs is Learning

One difficulty with Papineau's theory arises when attempting to extend his evolutionary explanations to include beliefs and desires. Unlike the selection of genes, the selection of beliefs is not strongly ancestral. I inherited my blond hair from my parents but not my belief that I have blond hair. Thus, while Papineau might be able to give an evolutionary account of my blond hair, he would be hard-pressed to do so for my beliefs.

Instead, Papineau shifts his focus to the selection of beliefs by the individual. Papineau offers what he terms an "anti-realist" account of belief. In contrast with the commonsense view which links beliefs to corresponding real facts in the world, Papineau claims that beliefs are "functionally identified states of natural beings" (xiii). In other words, beliefs do not need to represent objects as they really are; instead, beliefs represent objects as they are most usefully described for the believing system. Thus, to use Papineau's example, a person who chances upon a life-size papier-mâché oak tree will formulate a belief to the effect, "That is an oak tree." Even though the belief is not caused by a real oak tree, the belief that it is an oak tree is, in some sense, the only biologically proper response. If live oak trees and their papier-mâché counterparts are indistinguishable, then people will generally have a better chance of being correct by entertaining more complex beliefs like, "That is either an oak tree or a replica of an oak tree." However, Papineau correctly points out that there would be little biological use in adding such complexity to our beliefs because papier-mâché oak trees are not part of our normal experiences. According to realism, the latter, disjunctive belief is more accurate; however, the situation to which it corresponds does not have biological significance. Only if the world were suddenly populated by oak tree replicas would it be useful to leave our beliefs open to the possibility of indistinguishable replicas.

Thus the selection, for instance, of a belief about oak trees over a belief about oak trees or oak tree replicas occurs because only oak trees are part of our repertoire of normal causes. An oak tree replica would be highly abnormal. The historical aspect of representation enters with the definition of normality: "normal" beliefs are defined by what types of beliefs have proven to have advantageous behavioral effects in the past. I have never seen a life-size replica of an oak tree; therefore, I have no reason to suppose that the next oak tree I see might be a replica. In the same way that my evolutionary history has selected my blond hair, my experiential history has selected my belief that I see oak trees.

Papineau's description of his own theory in evolutionary terms is slightly misleading, since he ends up making representation depend upon both evolution and learning. For Papineau, "natural selection" can mean either the evolutionary selection process or the individual learning process: it "occurs within generations, by learning, as well as between generations, by genetic changes" (Papineau 66). Individuals select beliefs in the same way that species select genes: "We can think of learning as selecting components for our cognitive mechanisms, analogously to the way that inter-generational evolution selects genes" (Papineau 66). Representations are either learned or inherited genetically.

Evolution Provides the Wrong Kind of Explanation

Papineau's theory leaves unresolved the question of what, exactly, it means to inherit a belief or desire genetically. In fact, Fred Dretske argues in *Explaining Behavior* that the evolutionary aspect of Papineau's theory is unnecessary, that representation requires only a history of learning and not any kind of evolutionary history. According to Dretske, evolutionary accounts provide the wrong type of explanation for behavior. A representation derived solely from genetics would admittedly have a historical connection between the representation and its content; however, the connection would not be of the proper type. Inherited representations, if they were possible, would derive their representational power from the meaning of the system's ancestors, rather than from the system itself. But that is not how representations actually work: "Beliefs are internal states whose meaning explains the behavior of the systems *of which they are a part*" (Dretske 1990 829, Emphasis Dretske's). According to Dretske, Papineau's evolutionary theory of representation fails because it does not guarantee that a belief will mean something to the believer, rather than to the believer's ancestors.

Perhaps an analogy will clarify the argument. For Papineau, inherited beliefs acquire their representational power in the same way that an octopus's ink acquires its function. The ink's biological function depends upon its use by the octopus's ancestors as a defense mechanism, not upon the octopus's use of its own ink. Regardless of how, or even whether, the octopus uses its ink, the ink retains its original defensive function.

Dretske argues that this description of function, though accurate for the octopus's ink, does not carry over to representations. Consider as a candidate for an inherited representation the fear of falling — acrophobia. Jed, like most people, is afraid of falling. Given the option of falling twenty feet or not falling twenty feet, Jed would unhesitatingly choose the latter. However, unknown to Jed, his fear of falling exceeds that of most people: he was born with a pronounced acrophobia, one that causes trembling, weakness in the knees and dizziness whenever Jed finds himself at a great height. Luckily for Jed, he was born and raised in Iowa, where he has never had occasion to travel more than a few feet above the ground. Consequently, he is as yet unaware of his own acrophobia. However, tomorrow is Jed's fifteenth birthday, and his parents have promised to drive him to Chicago where they will visit the Sears Tower, the world's tallest building. Thus, tomorrow Jed will find out about his inherited fear.

According to Papineau's theory, Jed's inherited fear is a representation. Jed has a desire — the desire to avoid falling — that results from millions of years of natural selection. Humans share their fear of falling (along with, incidentally, ophiciophobia, the fear of snakes) with their primate ancestors who, living primarily in the trees, undoubtedly found such fears to be evolutionarily advantageous (Sagan 158). Jed himself has never had any use for such a fear. However, according to Papineau's theory the fear of falling has the proper selectional history to make it a representation. It has been naturally selected because of the biological advantage it provided for Jed's phylogeny, even though Jed himself, like an octopus in an aquarium, finds his inherited survival mechanism to be completely useless.

Jed raises a problem for Papineau's theory because, at least until tomorrow, Jed has a representation of which he is unaware. True, Jed knows that he doesn't like to fall from great heights, but that is a rational, reasoned fear which he has learned by watching television or simply by extrapolating from his own experience. What Jed doesn't (yet) know is that he has an extreme, instinctive, physical response to great heights, a response over which he has little or no conscious control. Tomorrow Jed will not learn to fear great heights; he will learn that he fears great heights. The phobia itself is pre-wired, built-in. As such, it seems on the level of a reflex, like a dog walking in circles to mat down the grass before lying down, even when the dog is only going to sleep on the living room carpet. The dog doesn't believe that grass is growing in the living room; walking in circles is an instinctive action which the dog's ancestors found useful in their life on the prairie. Similarly, Jed was not born with a built-in desire to avoid falling; fall avoidance has simply proved useful for many of Jed's ancestors. Nothing in his Iowan experience has given Jed reason to dread falling. Thus his "fear" of falling shouldn't be thought of as a fear at all: it is not a desire to avoid falling so much as a reflex which decreases Jed's chance of falling (thereby increasing his chance of surviving). Jed cannot inherit representations; he can only learn them.

The Connection Between Representations and Behavior (Dretske)

Dropping the evolutionary aspect of Papineau's theory leaves learning as the basis of representation. However, simply claiming that a system must learn in order to represent hardly seems profound. More needs to be said about what, exactly, is learning. Dretske offers a theory in which the process of learning provides the basis for representation. But Dretske's definition of "representation" also requires that the learning be of a particular type, that the representational system create links between internal, natural indicators and the external events which they indicate.

Definition of "Behavior"

Rather than starting with representation, as Papineau does, Dretske begins by defining "behavior." In order to distinguish an animal's behavior from the things that happen to it, Dretske points to the cause of the event: in order to qualify as behavior, an event must be caused by a factor internal to the behaving system. Thus, the simple event of a bee stinger penetrating a child's finger is insufficient to guarantee that the statement, "The bee stung the child," is true. After all, the stinging may have occurred as the child picked up a dead bee. "To get bee behavior, to have something the bee does, the cause of . . . stinger penetration must come from within the bee" (Dretske 1988 2).

This cause, in addition to being internal to the bee, must be of a particular type. Dretske distinguishes between two types of causes — "triggering" and "structuring." The former is what people normally mean by "cause" — the occurrence which triggered the event in question. The latter is a second-level type of causality: the structuring cause is that which causes the triggering event to cause the final event. For Dretske, behavior is a causal process (Dretske 1988 33); therefore, simply describing the triggering cause of the output is insufficient to explain behavior. In order to explain a causal process such as behavior, one must rely upon a second-level explanation in terms of the structuring cause — one must give the cause of the causing.

Natural Signs

As a step in his search for the structuring cause of behavior, Dretske looks to "natural signs." A natural sign indicates what it does by virtue of a phenomenological link to the environment. Thus, for instance, a rabbit's tracks indicate that a rabbit passed by. They do so by virtue of their historical link to the passing rabbit, rather than because of some arbitrary assignment, as the word "rabbit" is assigned to a family of long-eared, short-tailed mammals with long hind legs.

Natural signs are important to the theory of representation because they seem to be necessary for representation to occur. In general, in order for a system to have any chance at reliably representing something — of maintaining beliefs about X or desires for X — the system must have access to some physical element which indicates the presence of absence of X. In other words, the system must have access to natural signs. Without these indicators, a system could not hope to represent with any chance of representing truly. Unless Flakey has access to something that indicates the presence or absence of a rabbit, Flakey will have no way of determining the truth of the statement, "There is a rabbit nearby."

Furthermore, in order for a system to represent, the natural signs used by the system must be internal to the system. Dretske has already pointed out that, in order to differentiate between a system's behavior and the things that happen to it, the source of the system's behavior must be internal to the system. Therefore, a robot which acts because of what has been programmed into it from the outside does not actually "behave" in a genuine sense. The programmer cannot play the indicator role normally played by natural signs because the programmer acts upon the system as a separate agent, rather than as an internal part of the system.

Therefore, the above criticism of Lloyd's physicalism needs to be revised. There is one current physical characteristic of representational systems that is necessary: the system must contain internal natural signs. Even the most bizarre alien life form, if it were representational, would have to contain some means of gathering information from its environment. Without such indicators, the alien would be completely shut off from the objects it represents.

Definition of "Learning"

However, mere existence of such internal indicators is insufficient for representation; they must also be relevant to the explanation of behavior. Specifically, a system capable of representation must be able to use its internal indicators as guides through the environment. If a system is incapable of using its internal indicators as guides, then the indicators' content is wasted. Thus the system must be able to learn, where "learning" is used in a restricted sense to describe "the process in which internal indicators . . . are harnessed to output and thus become relevant — as representations, as *reasons* — to the explanation of the behavior of which they are a part" (Dretske 1988 104). In other words, the internal indicators in such a learning system are structuring causes for behavior. Here Papineau's concept of biological significance comes into play, for the harnessing of the system's output to its internal indicators means determining what is most biologically useful to the system under normal circumstances. Furthermore, the learning system fulfills Lloyd's original uptake condition for representation — the requirement that the content of a system's representations be causally relevant to the system itself — since learning involves developing a correlation between indicators and behavior.

The Completed Picture of Representation

Thus, instead of the three-part definition of representation proposed by Lloyd, Dretske offers a two-part definition. In order to represent, a system must

1. contain internal natural signs, and
2. have a history of learning in which the system's natural signs become linked to behavioral output because of what they indicate.

The first condition guarantees that the system have some connection to the content of its representations. Then, by means of the learning process, the system's interaction with its environment through these indicators becomes fitted to the qualities of that environment. In learning, the indicator properties of the system's internal elements cause a correlation to arise between states of those indicators and the appropriate behavioral responses. Because learning is a process, representations may be tuned to any degree of precision and accuracy, depending upon how much the system has learned. And since the process occurs solely as a result of the properties of the system's internal elements, the meaning of its indicators must be relevant to the system alone, and not to some external programmer.

Problem with Historical Definition: Cosmic Accident Replica

Unfortunately, the two-part definition of representation is not without its problems. Requiring representations to have a certain type of history leads the theory into trouble. Remember Papineau's little green bug that "coagulates by cosmic happenstance" (Papineau 74). The little green bug's limb might, believably, be functionless because it lacks a certain type of history. However, consider a similar example in the context of belief. Instead of a cosmic accident producing a bug, it produces a human being. The accidental person functions as well in human society as anyone else, yet, according to the historical definition of representation, the new person doesn't believe anything. The accidental person cannot believe anything because the person has not learned anything. The person has no representations because representation requires a history of linking natural signs to internal indicators.

Papineau's Response

The cosmic accident replica example provides a formidable stumbling block for any historically-based theory of representation. Papineau admits that the cosmic accident argument presents a problem for his teleological theory, but denies that the argument destroys his theory. He acknowledges that our intuitions lie in the other direction, that we would tend to grant a newly-created human being beliefs, as long as those beliefs are coherent and correspond reasonably well with reality. But he replies that "we ought to change our intuitions" (75). Papineau argues that our intuitions about, for instance, the function of a peacock's tail in attracting mates has been molded by Darwinian theory. Similarly, he claims, we should allow the teleological theory of representation to mold our intuitions about the cosmic accident argument.

Clearly, this argument by analogy cannot hold in general. It ignores the point of respecting naive intuitions. Darwinian theory molded our intuitions after it had been accepted on its own merits. If we begin altering intuitions to fit any new theory, then we eliminate one of our strongest discriminatory tools. While it may be true that the historical account is strong enough to change our intuitions about representation, the arguments in favor of a historical definition of representation would have to be strong enough for us to accept the theory first.

Historical Representation as Epiphenomenon

The cosmic accident replica example points to a deeper problem with any definition of representation based solely upon the history of the representing system: such definitions do not allow the property of representation to reside in the current system. To see this, consider a refinement of the cosmic accident argument. Instead of merely creating any human being, assume that the accident creates me, Bill Grundy. As a by-product, it accidentally vaporizes the old Bill Grundy. The change takes place instantaneously: one moment I'm me, the next moment, every atom in my body has been dispersed to outer space, only to be immediately replaced by an exactly similar structure of atoms. Let's assume that the accident occurred just as I sat down to write this paper. Has what I have written meant anything? According to Papineau, the answer is "No." Perhaps you the reader can interpret the words meaningfully, but I didn't mean it.

The replicated Bill Grundy is perfectly indistinguishable from the old Bill, except that one has the ability to represent and one does not. Talk to me. I don't believe anything I say. I don't really want any of the things that I claim to want. Even internally, I don't believe anything. Maybe I feel like I believe, but I can't even believe that I believe. Unfortunately, you have no way of knowing that I'm a replica. I can't even know I'm a replica. There is absolutely no criterion for distinguishing the representational from the non-representational. Representation is epiphenomenal.

This conclusion makes any historical theory of representation distinctly unsatisfying. The two-part definition of representation requires that a representational system have a certain type of history. But that requirement means that the inspection of any individual system, without knowledge of the system's past, cannot reveal whether the system has the ability to represent. However, if there is no phenomenological difference between representing systems and non-representing systems, then where is the point in arguing about them? If representation is merely epiphenomenal, then the attention paid it seems vastly beyond its due.

Representation as Function

One response to this dilemma turns representation into a function. In general, functions fundamentally depend upon the history of the functioning system. Therefore, if representation can be proven to be a function then its existence should depend upon the history of the system.

The argument for representation as function runs by way of analogy. Consider a metal object dug out of the ground by anthropologists at the site of an ancient village. The object might be extremely spoon-like in appearance. In fact, it could be physically identical to the modern spoons the anthropologists eat with each day. However, a physical description of the metal object is insufficient to guarantee that the object is, in fact, a spoon. The anthropologists need to know the history of the object before they can accurately describe its function. In order to determine the object's function, the anthropologists must know whether the object was used, or at least intended for use, as a spoon. It may, after all, have actually been used as a very small shovel. Its function depends upon its history.

Similarly for representation, the history of the system determines its ability to represent. The relevant history is learning-based: the occurrence of beliefs and desires depends upon the individual's learning history. Without this requisite history, the system cannot represent. Just as two identical spoon-like objects can be a spoon and a shovel, so two identical people can be a normal, representing human and a non-representing cosmic accident replica. This description of representation as a function makes sense of representation's dependency upon history.

Wright's Definition of "Function"

However, simply stating that representation is a function will not solve the problem. The argument requires a definition of "function" to which representation may be compared. Larry Wright provides such a definition in "Functions":

> The function of X is Z *means* (a) X is there because it does Z, (b) Z is a consequence (or result) of X's being there. (363)

The two halves of this definition point out two distinct characteristics of functions. Part (a) requires that the function be relevant to the history of the system. For living organisms, this requirement means that the existence of the functioning part of the system must be explicable in terms of natural selection and specifically in terms of its ability to do Z (whatever that may be). For example, the function of the epiglottis is to cover the space between the vocal chords when swallowing. Presumably, the epiglottis's ability to cover that region accounts for its evolutionary development: if the epiglottis lacked the ability to cover the glottis, then it would not be in the back of our throats. The function of the epiglottis explains why it is there.

The second half of Wright's definition concerns the causal efficacy of the functioning part of the system. "Z is a consequence (or result) of X's being there" means that Z follows, in a loose sense, from X. Most importantly, it means that X does not follow from Z. Consequence is an asymmetric relation. For the epiglottis, the second half of Wright's definition translates to, "The glottis's being covered is a consequence of the epiglottis's being there." This makes sense, since if the epiglottis weren't there, the glottis simply would not get covered. And just as clearly, the relation is asymmetric: the existence of the epiglottis does not follow directly from its ability to cover the space between the vocal chords. The type of consequence Wright requires for this second half of the definition differs sharply from the indirect causal connection of part (a). In Dretske's terms, the first half of Wright's definition calls for a structuring cause — it asks why X causes Z, why we happen to be constructed so that the epiglottis has the ability to cover the glottis. On the other hand, the second half of the definition requires that the functioning part of the system be a triggering cause, that the epiglottis actually cause the glottis to be covered.[2]

Representation and Wright's Definition

The argument for representation as a function relies only upon the first half of Wright's definition. The ancient, spoon-like object has the function of a shovel because it was created for its shovelling abilities. The object is there because it can shovel.[3] Similarly, according to the learning theory of representation, beliefs and desires exist because of their function as survival mechanisms. They are chosen in a learning process akin to natural selection in which the system develops representations which correspond more and more closely with what is most useful to the system. Thus representations are there because they contribute to the survival capacity of the organism.

The second half of Wright's definition requires that the organism's ability to survive be a consequence of representation. The internal natural signs, by linking semantic content directly to behavior, provide this causal relation between representation and survival. Essentially, part (b) of Wright's definition points out that, in order to be functional, representations must be able to do something. An artifact that looks like a golf ball clearly could not have the function of a spoon because it lacks the ability to scoop up food. Without some type of ability, representations cannot possibly have a function. Simply having a certain type of history cannot possibly be sufficient, since such a property, which has no physical instantiation in the system now, cannot be relevant to the system's survival. That is, in a sense, why representation requires not only a history but a history that involves internal indicators, elements within the representational system which provide its beliefs and desires with causal efficacy. Thus the learning history theory of representation fulfills both halves of Wright's definition of function: (a) representations are there because they aid survival and (b) the system's survival is a consequence (or result) of the representations' being there.

And since representations are functions, the cosmic accident replica argument against historical definitions of representation loses its edge. In the same way that the "spoon" unearthed by the anthropologists is actually not a spoon at all, so the "representations" of a cosmic accident replica are not representations at all. Since an object's function depends in part upon the object's history, a replica which lacks any history at all cannot possibly have a function. Functional objects cannot randomly coagulate. Therefore, the cosmic accident replica provides no challenge to the historically-based learning theory of representation.

[2]Actually, Wright's consequence is not that strict. The consequence relation can hold even though the result never occurs: the function of the epiglottis remains the same even in someone whose diminutive epiglottis fails to do its job.

[3]This may be a weakness in Wright's definition. Cummins argues that, at least for conscious functions, an artifact's function may depend upon its use, rather than simply the creator's intent. A zodiac mosaic with a fallen piece of ceiling tile embedded in its center could function as a sundial. It could even have the function of a sundial, argues Cummins, if the local people have used it as a sundial for many centuries. However, my argument does not depend upon such a distinction.

Representational Machines

Given this theory, the question that remains unanswered is the one this paper began with: "What about Flakey the Robot? Can Flakey represent?" The answer, it seems, is that while Flakey cannot now represent, the robot might learn to represent. What is required for a representing machine is a learning ability of the kind described by Dretske. This would require that the machine have internal indicators which are linked to the environment and which configure themselves according to the constraints imposed by the environment. Creating such a machine is, in principle, possible.

The first requirement, internal indicators, can be fulfilled by any type of transducer. Even Flakey's simple sonar sensors qualify. Based upon the same principles as bat sonar, these transducers convert information about the distance between the sonar and nearby objects into electrical pulses in the robot. Such pulses are, in Dretske's terms, natural signs which indicate distance information.

Machine Learning

The challenging task, then, is allowing the robot to learn. Unfortunately, the most simple approaches to learning will not satisfy Dretske's definition. Consider Flakey's wall-following algorithm. Flakey uses information from its sonars to keep track of the distance between the robot and the wall, attempting to maintain that distance within a certain range of values. If the wall gets too far away, Flakey turns toward the wall; if the wall comes too close, the robot turns away. The wall-following program necessarily started out with a number of constants, such as the minimum and maximum distances allowed, the rate and degree of the compensating turns, and so on. And initially I had to set these parameters somewhat arbitrarily. Later, after running the robot and watching the results, I was able to choose better values for the constants. However, trial and error is not the only possible approach to setting parameters. Rather than learning about the parameters myself, I could have set the parameters by instantiating a learning algorithm in the robot. The learning program could, for instance, keep track of the number of times Flakey hit the wall and the number of times the robot lost track of the wall. By altering its internal parameters in response to the number-hit and number-lost records, the robot could gradually improve its ability to follow walls. Flakey would learn how to follow walls better.

Habituation vs. Learning

However, such learning hardly seems profound. In Dretske's terms, a system which adjusts its response as a result of repetitive stimulus does not engage in true learning, but merely in "sensitization" or "habituation" (Dretske 1988 96). And certainly in the case of Flakey's wall-following routines, the so-called learning seems to be of a different type from human learning: the robot can change its parameters, but not its approach. The basic algorithm must remain the same.

Beyond Habituation

Fortunately, habituation is not the limit of robotic learning abilities. A meta-heuristic approach would allow Flakey to explore a wider range of possibilities: if the robot had a means of choosing between various heuristics, then its range of behavior would increase exponentially. In search of improved heuristics, the robot could alter its own program (while keeping a copy of the previous version, just in case of errors) to find more beneficial modes of behavior. Alternatively, Flakey might rely upon a parallel distributed network. These networks closely model, at least behaviorally, the learning abilities of the human brain.

They therefore seem like a likely candidate for instantiation in a representational robot (Humans are, after all, presumably representational beings.).

Regardless of which approach to learning is chosen, the important point is that robots can learn. That robots do not yet learn as well as humans results not from a principled difference between the physical functioning of the brain and the limitations of semi-conductors, but from our own lack of understanding of the brain and from its sheer complexity. A complete model of the human brain might require a hundred years to build and end up the size of the Sears Tower, but it is, in principle anyway, a possibility. Certainly a mobile robot with the ability to learn in a suitably plastic manner is plausible.

Duplication of Representation

But the significance of such a robot is unclear. Beyond its practical importance (which would be great), Dretske implies that if Flakey were capable of using information from its transducers to significantly modify its internal architecture, then the robot would be a representational system. However, there is a refined version of the cosmic accident replica problem which calls into question the significance of Flakey's representations. Presumably, after Flakey's self-modification process had generated one representational robot, its creators would attempt to create more of the same. But in mass-producing their new intelligent robot, the creators would not start each one from the same rudimentary beginning. They would turn Flakey off, open it up, and look inside. And they would not see anything profound about this particular robot. Recording the specifications of the self-modified Flakey and producing a thousand exact replicas would result in a thousand and one robots, all with the same behavioral potential. According to the learning theory of representation, however, only the original would be representational, since only the original would have gone through the process of harnessing its internal indicators to behavioral output. This example revives a problem first raised by the cosmic accident replica: differentiating between physically identical robots — calling the original representational and its replica non-representational — seems counter-intuitive.

Derived Representation

The simplest solution to this dilemma gives representational power to both the original Flakey and its replica, Flakey II. According to this argument, the duplicate robot maintains the necessary historical connection between its internal indicators and what they indicate by means of the person who carries out the duplication. A cosmic accident replica comes into being randomly. In contrast, a duplicate of Flakey the Robot comes about because of the intent of the programmer. The randomly generated system clearly lacks the proper learning history; however, an intentionally replicated system does seem to have access to such a history. The Flakey replica has all the learning history of the original Flakey plus one more link — the intentional transference of representation through the person who carries out the duplication. This sort of derived representation, while indirect, might be on a par with genuine representation.

If the two types of representation — original and derived — are equivalent, then the whole project of making a learning robot seems superfluous. After all, if humans can transfer the power of representation from Flakey the Robot to Flakey II, then why not transfer that power directly from themselves to robots? The human programmer has undoubtedly already learned about the impossibility of walking through walls. Therefore, making Flakey develop its own obstacle avoidance routines simply in order to learn the same principle seems redundant. The programmer's learning history, rather than the robot's learning history, can serve as the historical basis for representation.

Meaning of Derived Representations

Of course, the problem with derived representation has already been pointed out by John Searle. Although these derived representations may have meaning for the system that learned them, the process of duplication does not transfer that meaning. Like the subject in the Chinese room, Flakey II receives a system of inputs and outputs which, to the robot, are entirely meaningless. In Dretske's terms, the derived representations should not be called "representations" at all because their semantic content does not contribute to the explanation of behavior. The explanation of Flakey II's behavior lies outside the robot — either in the original, intelligent prototype or in the creators who carried out the duplication of the prototype. The derived representations are similar to Jed's fear of falling. Jed's acrophobia causes him to tremble and sweat while on the sky-deck in the Sears Tower. He appears to harbor an actual fear of heights, when in fact he has only a reflexive, physiological response to high places, a response that doesn't involve his conscious brain functioning at all. Similarly, the behavioral responses built into Flakey II are reflexive and non-representational. Only when Flakey II begins its own learning process will the duplicate robot develop the capacity to represent.

Summary

Thus the learning theory of representation provides a picture of Flakey's representational capacities as shown in Figure 1. Each of the four circles in the figure represents one version of Flakey. #1 is the learned Flakey, whose history of learning stretches to the left in a dark line. The second Flakey is the cosmic accident replica, whose only history consists of a burst of cosmic rays. Flakey #3 is an exact duplicate of the learned Flakey. The duplication process, represented by the thin, curved line, is mediated by a scientist (the "S" in the square). Of course, the scientist has her own history of learning, so Flakey #4 is a robot created directly by the scientist herself. According to the theory of representation outlined above, the original Flakey can represent by virtue of its history of learning, of coordinating the states of its internal indicators with appropriate behavioral responses. However, none of the other robots — numbers 2, 3 and 4 — can represent because none of them has a learning history.

Because of the argument for representation as function, Flakey #2, the cosmic accident replica, does not cause a problem for the learning theory of representation. Just as the function of a spoon depends upon the spoon's history, so the ability of a system to represent may also depend upon the system's history. Since Flakey #2 maintains no historical ties between its own internal indicators and what those indicators indicate, the robot has no access to genuine meaning. The cosmic accident replica has no history; hence, it cannot represent.

Unfortunately, the functional argument cannot explain why Flakey #3 does not represent. The difference between #2 and #3 lies in the means of replication: the cosmic accident replica occurs entirely by chance, whereas Flakey #3 is created intentionally, by a scientist who wants to duplicate the original, representational Flakey. This difference removes the force of the function analogy, because an intentional duplication of a functional item does result in the transfer of the function. If #1 were a spoon, and #3 were a spoon-like object created by a scientist who intended to create a spoon exactly like #1, then #3 would actually be a spoon. In fact, spoon #1 is not even necessary; the scientist could just create spoon #4 without modelling it on any prior spoon at all. So functions do transfer through intentional replication, that is, across the thin, curved lines in the diagram above. The argument that representations are historical in the same way that functions are historical fails to differentiate between the representational Flakey #1 and the non-representational Flakey #3.

However, Flakey #3 must be non-representational. Otherwise the theory violates the point made by Searle's Chinese Room example. Searle shows that Flakey #4, although intended by its creator to have a certain set of meanings attached to its inputs and outputs, cannot possibly have its own beliefs and desires.

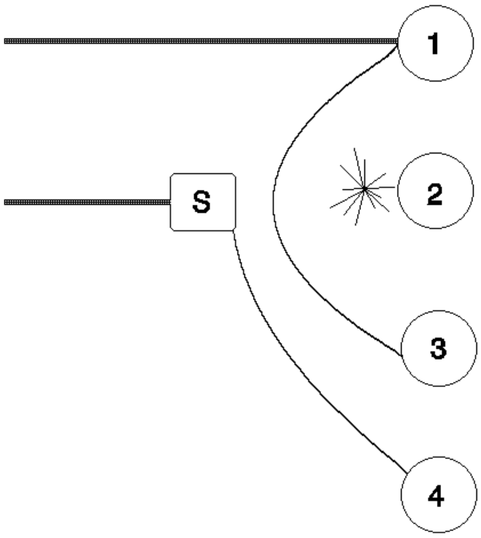

Figure 1: Four versions of Flakey. Only #1 represents.

The content of its "beliefs," as described by the scientist, are completely inaccessible to the robot. Just as the scientist cannot attach her own learning history to Flakey #4, she cannot attach the history of Flakey #1 to Flakey #3 simply by duplicating the first Flakey. Even though the correlations between internal indicators and behavioral output have representational meaning for Flakey #1 and for the scientist, for Flakey #3 the correlations are just so many reflexive behaviors. Their semantic content is completely irrelevant to explaining the behavior of the replicated Flakey. Hence, neither Flakey #3 nor Flakey #4 can represent.

A Final Test

But does this theory, in which only a robot that does its own learning has the ability to represent, make sense? Here's one final test: switch the four circles in the diagram above from robots to people. #1 is you; #2 is an exact replica of you coagulated by chance from cosmic rays; #3 is your clone; and #4 is a person who by chance resembles you in every way, but who was created by some super-scientist in another galaxy. The arguments outlined above should still apply to this human example: only you — the original you — can represent. The other versions have no beliefs or desires, despite their ability to mimic your every behavior.

At this point, intuitions begin to falter. On the one hand, some people hope that human clones could not think in the same way that we do. There is something seemingly unnatural in supposing that a replica of yourself could have all of your beliefs and desires. On the other hand, an atom-for-atom replica of yourself would be entirely indistinguishable, not just physically but also behaviorally, from the original. And for everyday experience, behavior serves to differentiate between representational systems and non-representational ones, at least for judgements between individuals.

But the problem of how we, as humans, differentiate between humans and clones is an epistemological one. Without even approaching that problem, however, the learning theory of representation generates problems of its own. Consider any replica of myself. The moment the replica comes into being it has no representations. For memories, this conclusion makes a certain amount of sense: the replica may believe that he attended kindergarten in 1973, even though, in fact, the replica didn't exist until 1991. However, the theory doesn't say that replicas come into being with a false set of beliefs; the theory claims that replicas don't have beliefs at all. The replica doesn't believe anything about kindergarten, although he does have the ability to learn about kindergarten. Unfortunately, that learning ability sets the replica up for a multitude of false beliefs about himself. Let the replica exist for a few years. Give him time to learn a language. Then ask him, “Did you attend kindergarten?” We can’t know what happens inside the replica, that is, whether he really thinks or not. But we do know that he will say “Yes” (He is, after all, behaviorally indistinguishable from me). And after the replica has heard his own reply, he can learn from that reply. He might now believe that he attended kindergarten. But that belief is false. At the very least, the replica believes that he believes he attended kindergarten. But that too is false.

False beliefs alone aren’t sufficient to destroy the learning theory of representation; however, the problems posed by the human replica do give cause to examine our philosophical project from a wider perspective. This paper started out trying to determine whether robots might be able to think, but that proved to be too difficult to determine. Instead, we focused on whether robots could be intentional, or representational. But nothing guarantees that intentionality is sufficient for human-style thought. True, we’ve never seen anything that could represent but which lacked the full range of mental functioning, but we’ve also never seen a robot capable of learning, in Dretske’s sense of the word.

Qualia

Perhaps there is more to the mind than mere representation. Our theory of representation does not account for “qualia,” the subjective character of experience. Just as the experience of the color red has a certain character which cannot be communicated other than by saying “red,” so the experience of believing, for instance, that the apple is red has a feeling associated with it. The learning theory of representation shows how to make Flakey represent, not how to make give Flakey the qualia associated with representation.

This distinction between representation and qualia may clarify our clouded intuitions regarding non-representational human replicas. Most of our naive intuitions regarding representation find their basis in the subjective character of beliefs and desires, rather than in a theoretic understanding of representation. We imagine what beliefs seem like cognitively to ourselves, and then we try to determine whether a replica of ourselves will have the same kind of experience. The discussion above, regarding my own replica’s beliefs about kindergarten, attempts to avoid qualia entirely by describing the replica as saying, “Yes, I went to kindergarten.” This public announcement allows us to avoid physically delving into the replica. But the replica’s reporting of its own beliefs serves in the example as a means for us to gauge his internal state. The naive interpreter needs to hear the replica say, “I went to kindergarten,” so that the interpreter will know that the replica believes — in the everyday, subjective sense — that he went to kindergarten. Thus when we ask ourselves whether a replica could think, we first consider what it feels like to think and then attempt to determine whether that feeling could be duplicated in a replica. Since the learning theory of representation is not a theory of qualia at all, the naive intuition that replicas could duplicate the qualia that we normally associate with beliefs and desires does not pose a challenge to the theory. In order to sort out our intuitions about replicas, we would need a theory of qualia, rather than a theory of representation.

Conclusion

But a theory of qualia will have to be expounded elsewhere. For now, we will be content with the learning theory of representation and its implications. According to our theory of representation, in order for Flakey the Robot to be able to represent, the robot must go through a specific type of learning process. In this process, as internal indicators in the robot gradually become aligned with appropriate behavioral responses, Flakey would develop a set of representations. Equating these representations with our normal beliefs and desires leads to difficulties with respect to replication: the person on the street who imagines a replica typically imagines a replica with qualia like our own. But having qualia doesn’t necessarily entail having representations. The intuition that a replica has a belief is really an intuition that the replica feels like it has a belief. Hence the difficulty with replication examples does not arise from the learning theory of representation; it arises from the implicit equation of representations with belief-desire experiences. By denying this equation and dividing qualia from representations, we allow the learning Flakey to have representations without implying that it must have the accompanying qualia. Similarly, the theory indicates that replicated Flakeys cannot represent; it does not claim that replicated Flakeys cannot harbor the kind of qualia that usually accompany representations. The question of which of the Flakeys has qualia must remain unanswered here. Thus, the learning Flakey may or may not think in the same way that we do; however, as a result of its learning capacity, the improved Flakey certainly does have representational power.

Works Cited

Bechtel, William. *Philosophy of Mind: An Overview for Cognitive Science.* Hillsdale, NJ: Lawrence Erlbaum, 1988.

Chisholm, Roderick. “Intentionality.” *The Encyclopedia of Philosophy.* Ed. Paul Edwards. New York: Macmillan, 1967.

Dennett, Daniel. *Content and Consciousness.* New York: Humanities Press, 1969.

Dretske, Fred. *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: MIT Press, 1988.

Dretske, Fred. “Reply to Reviewers.” *Philosophy and Phenomenological Research.* L.4 (1990): 826–31.

Fodor, Jerry. “Semantics, Wisconsin Style.” *Rerepresentation: Readings in the Philosophy of Mental Representation.* Ed. Stuart Silvers. Boston: Kluwer Academic, 1989.

Lloyd, Dan. *Simple Minds.* Cambridge, MA: MIT Press, 1989.

Lopez, Barry. *Arctic Dreams: Imagination and Desire in a Northern Landscape.* Toronto: Bantam, 1986.

Papineau, David. *Reality and Representation.* Oxford: Blackwell, 1987.

Sagan, Carl. *The Dragons of Eden.* New York: Ballantine, 1977.

Searle, John. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences.* 3 (1980): 417–58.

Wright, Larry. “Functions.” *Conceptual Issues in Evolutionary Biology.* Ed. Elliot Sober. Cambridge, MA: MIT Press, 1984.