# Functionalism and Artificial Intelligence

Kevin Connor

One of the most potentially important and least successful projects in computing during the previous half-century has been human-level artificial intelligence.[1] This project has been daunting not because we do not understand the capabilities of the technology; indeed it has been proven that all computers, though some may be faster or more efficient than others, are nevertheless no more powerful in terms of the sorts of things they can compute than Alan Turing's abstract notion of a computing device, which he first presented in 1936(Stanford). Instead, the difficulty in programming artificially intelligent machines stems in large part from our lack of understanding of how our own minds function. AI research has tended to follow our uncertain and certainly unproven models of human intelligence-and hence its failures have tended to rest upon the failures of these models. My purpose here is to illustrate the decisive failure of one of the most important of these models, and present an alternative view that offers hope for the ultimate viability of strong AI.

In *Representation and Reality*, Hilary Putnam attacks what is for computer scientists working on AI the most promising theory of human intelligence, functionalism. Putnam's definition of the functionalist model says that "…psychological states ('believing that *p*,' 'desiring that *p*,' 'considering whether *p*,' etc.) are simply 'computational states' of the brain. The proper way to think of the brain is as a digital computer. Our psychology is to be described as the software of this computer-its 'functional organization.'" (Putnam 73) Putnam's arguments in *Representation and Reality* are convincing, and I will discuss them in the first portion of this paper. In the latter portion, I will present an idea from Roger Penrose: that the promise of a model of the mind that is based in quantum physics may allow us to rethink the nature of the relationship between computers and the mind, providing a new research paradigm for strong AI in a world without the functionalist model.

The simplest version of functionalism that Putnam describes is known as the *Single Computational State* functionalism. In this con-

ception, each possible propositional attitude is describable in terms of a single state, which remains static across all physically possible organisms ("physically possible organisms" include machines). That is, "believing that snow is white" is supposed to be the same computational state for all organisms capable of having that belief." (Putnam 80) Putnam then conceives the sort of model that is necessary for an organism to function in this manner. Obviously, some sort of thinking language or "mentalese" is necessary, and also some kind of function which determines whether new sentences are sufficiently understood to be added to the language ("c-function" for Putnam). This organism will also require what Putnam calls a "rational preference function"(Putnam 80) in order to decide how to act in any given situation, together with the c-function described above. The rational preference function would need some variables to mark the particular desires of the organism (i.e., when it is raining outside, to allow for the possibility that I'm sad because I want to play baseball, or that I'm happy because I'm a farmer).

To illustrate how this conception would work, suppose that I am such an organism and I am presented with a new propositional state, "kittens are small and fuzzy", and let us further suppose that this is a perfectly adequate definition of the essence of "kitten": that all organisms who had a complete understanding of kittens agreed that they are best described as small and fuzzy. In order to process this state, my c-function would check my degree of understanding of "small" and "fuzzy" in order to see whether I know enough about the component parts of the propositional state in order to allow it into my thinking language. If I allow the attitude into my language, then I assign it a degree of understanding based on my degrees of understanding of the component parts, and can now access thoughts and judgments about kittens with the aid of my rational preference function

What Putnam finds troubling is that when we attempt to figure out meaning in this model, "all we are given to go on is the current subjective probability metric (the current degrees on confirmation), the current desires (the current "utilities"), and

the underlying *c*--function by which the current subjective probability metric was formed on the basis of experience." (Putnam 80) He says that at least the first two of these things might be totally different, even for meanings that we would like to say are the same, in different organisms: organisms will undoubtedly have different degrees of understanding of different sentences in the mental language, and will undoubtedly feel slightly different desires about those words. The result of this is that when you and I say "kitten," we can never mean the same thing, which is undoubtedly unworkable.

The problem is not solved even if we assume that there are sentences ("kittens are small and fuzzy" might be one of them) that are *analytic* terms that are universal across all organisms. Putnam says that this isn't going to work because we couldn't say that "small" and "fuzzy" have the same meaning for us analytically, because Putnam has shown that meanings cannot exist solely in the mind-that there is a linguistic division of meaning for words.[2] If these words don't have analytic meaning for us, then it seems that we can't come to the same analytic meaning for the whole propositional state "kittens are small and fuzzy," or further any propositional state. As Putnam sums up this line of reasoning, "there is no way to identify a computational state that is the same whenever any two people believe that there are a lot of cats in the neighborhood (or whatever). Even if the two people happen to speak the same language, they may have different stereotypes of a cat, different beliefs about the nature of cats, and so on (imagine two ancient Egyptians, one of whom believes cats are divine while the other does not)." (Putnam 82).

Another form of functional formalism that he briefly considers he calls *sociofunctionalism*: "Why not think of the entire society of organisms together with an appropriate part of its physical environment as analogous to a computer, and seek to describe functional relations within this larger system?" (Putnam 74) For example, the state of "thinking that there are a lot of cats in the neighborhood" may be describable in terms of the thoughts that each person in the neighborhood has about the cats

-each of their individual states builds together the full functional state quoted above. This is obviously a complicating move: we will have to draw functional relations across many different types of organisms and environments to create the full functionalist picture, which would be perhaps in principle possible.

Putnam in short says that defining this complete system is a pipe dream. He says that when different people speak the same word, they inevitably have at least slightly different mental conceptions of that word, like the cat example from our discussion above about simple functionalism. We need some way of arbitrating these meanings; of deciding whether a particular conception fits a criterion that he calls "reasonableness." In a society of millions of people, each with her own definition of cat, there must be some way of deciding which definitions are more correct or more complete; and the "real" definition would be synthesis of those that are most "reasonable." He explains, "…this, I have argued, would be no easier to do than to survey human nature *in toto*. The idea of actually constructing such a definition of synonymy or coreferentiality is totally utopian." (Putnam 75)

That is, the project would involve a listing of uncountable (in the mathematical sense of infinite) possibilities of definitions in uncountable languages-there is no reducible formula for "reasonableness." Putnam concedes that such a system may be in principle possible, noting that "Few philosophers are afraid of being utopian…"(Putnam 76) But my purpose here involves the implications of the demise of functionalism for artificial intelligence, and we need a reducible formula to program machines. The question of whether such a listing is in principle possible is moot to the AI programmers.

Putnam supposes yet another way to reconceive the functionalist argument. This argument shifts the burden from computational *states* to computational *relations*, specifically *equivalence relations*. For example, we could try to figure out if when I say the word "cat" in my particular environment X and a Thai speaker says the word "meew" (which means "cat" in Thai) in her particular environment Y, whether we are in fact talking about the same concept. If we can enumerate all of the physical

details involved the Thai conception of meew and the English conception of cat-admittedly a difficult project-then we can create an equivalence relation of the form "cat as used in English in this particular situation X is synonymous with meew as used in Thai in this particular situation Y." As Putnam explains, this relation (and ones like it) "…must be a predicate that a Turing machine can employ: a recursive predicate or at worst a "trial and error" predicate." (Putnam 85) Since all computers have been proven to be as powerful as Turing machines (and therefore each other) and recursive and "trial and error" (which we might call exponential) algorithms *are* computable, if slow, this gives great hope to a functionalist model of artificial intelligence.

Obviously, this argument rests on the assumption that our minds function in a way very similar to Turing machines. But Putnam does not need to refute that claim to make his objections. First he notes that in order to make the difficult decision mentioned above regarding whether "cat" and "meew" actually refer to the same extension, we have to know a whole lot about the linguistic and environmental conventions in the situations. Without careful consideration of how the Thai language is used, it might appear that meew refers only to "Siamese cat" (it in fact refers to all cats) as those are the only sorts of cats that one encounters in Thailand. And there are uncountable variables like this that need to be considered, even just for our example. As Putnam says, "What is at stake…is the interpretation of the two discourses as wholes." (Putnam 86) In any given discourse, it is necessary to learn something of the discourse before we can understand its terms. One cannot know what "existentialism" means without knowing some philosophy, or what "adverb" means without knowing some English, for example. Putnam finds two critical problems relating to this idea that occur for this theory.

Putnam imagines a situation in which there are two scientific theories from two different cultures, one from Mars and one from Venus. These theories are about the same phenomenon, and are so similar that an outside observer would regard their meanings as identical, once he had discerned that their environ-

ments were such that particular terms in the theory had identical meanings. If we are to analyze how we would make this claim about particular terms in the theories, we need to answer the question of what each term actually refers to in each culture, which will likely involve determining whether the theories are true for each culture. But if the theories are about large enough cosmological concepts, we need to know information about the whole universe before we can judge whether the theories are true, which is an awfully large reference set. As Putnam concludes, "…the assumption that in principle one can tell what is being referred to by a term used in an environment from a sufficiently complete description of that environment in terms of some standardized set of physical and computational parameters is false *unless we widen the notion of the speaker's environment to include the entire physical universe.*" (Putnam 87) Considering the entire universe is obviously going to make the problem incomputable, which will bring down this theory of functionalism as far as AI is concerned.

The second problem is related: "any theory that 'defines' coreferentiality and synonymy must, in some way, survey all possible theories" (Putnam 87) For example, there are many different theories of functionalism, some of which we've looked at thus far in this paper and some of which we haven't. If we are ever to write the equivalence relation that defines how we can tell whether a particular functionalist theory or element of a functionalist theory is then synonymous with another, we have to consider not only all possible theories of functionalism in existence, but also *all possible* functionalist theories. The trouble with this is that human societies are by their nature progressive in terms of how they conceptualize the world, so it is unclear how a human being living in any given society could account for these theories that have yet to be invented. As Putnam says, "To ask a human being in a time-bound culture to survey all modes of human linguistic existence-including those that will transcend his own-is to ask for an impossible Archimedean point." (Putnam 89)

Putnam's defeat of these and other forms of functional-

ism and his conclusion that our minds are not in any way reducible to machine states or functional languages seems convincing to me. Since these computable states are precisely what are necessary for traditionalist conceptions of AI, it seems that the downfall of functionalism is a deadly blow for the future of AI research. A new research paradigm for strong AI is sorely needed. In *Shadows of the Mind,* Roger Penrose presents how such a paradigm can be conceived.

Penrose explores how the young science of quantum mechanics might be brought to bear on our conceptions of consciousness. Quantum mechanics is a reduction of empirical realities to probabilities. When we attempt to understand subatomic particles, it turns out that we can't say as much as classical Newtonian physics says we ought to be able to about each individual particle. Indeed, basic facts about particles like position and velocity are inevitably altered depending on the sorts of measurements we take. In short, quantum mechanics can tell us probabilistically how a large number of particles will behave in a certain situation, but cannot ever predict for any given particle precisely what that particle will do.

Penrose says that scientists have been generally unwilling to consider modeling the mind on a large scale using quantum mechanics. They might admit that on a small scale there may be quantum interactions taking place between the atoms of the brain, but "…it seems to be generally assumed that it is quite adequate to model the behavior of neurons themselves, and their relationships with one another, in a completely classical way. "Penrose 348) Once we have committed ourselves to modeling small parts of a system in a Newtonian fashion, accepted scientific practice necessitates that we model the system itself in a classical way as well. The result is a standard Newtonian model of brain function.

Penrose argues that it may be possible to define a theory of the mind as a whole that is based on quantum physics-that the whole brain itself could be described as an example of *quantum coherence*, which refers to "…circumstances when large numbers of particles can collectively cooperate in a single quantum state

which remains essentially unentangled with its environment." (Penrose 351) This coherence would allow particle-level quantum interactions to have an effect on a system as large and complex as the brain. If quantum coherence could be demonstrated, it could act as a bridge between the concepts of brain and mind. The chemical and physical functioning of neurons in the brain could follow a Newtonian model and the functioning of the mind as we experience it could be explained by quantum coherence.

How might the mind exhibit this quantum coherence, then? Penrose notes that our understanding of the brain has led to a classical picture "…in which neurons and their connecting synapses seem to play a role essentially similar to those of transistors and wires (printed circuits) in the electronic computers of today." (Penrose 352) Given this understanding, we can and must use a classic computational model for this part of the structure. However, Penrose also says that research shows that the strength of these connections and even the physical connections themselves change over time-almost as though the silicon and steel in your personal computer were to rearrange themselves on a regular basis. The classical model attempts to explain this computationally, but Penrose finds as Putnam has that computational models inevitably fail at explaining such behavior. Penrose concludes, "…we must look for something different, as the appropriate type of controlling 'mechanism'-at least in the case of synaptic changes that might have some relevance to actual *conscious* activity." (Penrose 354) Large-scale quantum coherence in the brain between individual neurons is a promising candidate for that mechanism. Furthermore, the young science of quantum computing, which uses principles of quantum mechanics and classical computing together to store data and perform operations simultaneously and flexibly on many particles, may eventually produce machines efficient enough to simulate this coherence.[3]

Penrose admits that the obstacles to constructing a quantum theory of the mind are large. He says, '[a human-level device] would have to incorporate the same kind of physical action

that is responsible for evoking our own awareness. Since we do not yet have any physical theory of that action, it is certainly premature to speculate on when or whether such a putative device might be constructed." (Penrose 393) The difficulty arises because once we have our physical theory, we'll also need a corresponding breakthrough in psychology that explains the connection between the quantum model and consciousness. This might seem to be a dubious proposition-Penrose admits that he has no idea how it might come about. However, it is in principle possible to model the mind in a quantum fashion, while Putnam has decisively ruled out modeling the mind in a classically functionalist way.

If we could come up with these theories, we could then construct a machine whose physical states corresponded to the way physical states work in our minds, and would be functionally equivalent to humans. Although the prospect of truly artificially intelligent machines looks grim in the near term, we should not yet give up hope. Penrose says, "…in a clear sense, these are still early days in the physical understanding of our universe-particularly in relation to mental phenomena." (Penrose 393-4) However, two and three bit quantum computers have already been built, which are capable of data sorting and simple arithmetic. As brilliant a mind as Richard Feynman believes that advances in quantum computing will stimulate advances in quantum physics.[4] What Penrose has offered us is a research paradigm: strong AI researchers who have been treading water with functionalism can turn to a quantum model and begin solving these difficult problems.

## Notes

[1] Or "strong AI;" see Searle, "Minds, Brains, and Programs"(1980), also Dreyfuss, What Computers Still Can't Do (1992)

[2] See the Twin Earth examples in Representation and Reality; also Putnam's article *Meaning and Reference*

[3] See www.qubit.org; Arrighi, P. "Quantum Computation Explained to My Mother," *EATCS* June 2003; Steane, A.M. "Quantum Computing," *Reports on Progress in Physics* vol. 61 (1998)

[4] See www.cs.caltech.edu/~westside/quantum-intro.html

## Bibliography

Penrose, Roger. *Shadows of the Mind*. Oxford University Press: Oxford, 1994.

Putnam, Hilary. *Representation and Reality*. The MIT Press: Cambridge MA, 1988

The Stanford Encyclopedia of Philosophy, "Turing Machine." http://plato.stanford.edu/entries/turing-machine/