# PUTNAM'S PROBLEM OF THE ROBOT AND EXTENDED MINDS

JACOB BERK

10.33043/S.15.1.88-99


## ABSTRACT

In this paper, I consider Hilary Putnam's argument for the *prima facie* acceptance of robotic consciousness as deserving the status of mind. I argue that such an extension of consciousness renders the category fundamentally unintelligible, and we should instead understand robots as integral products of an extended human consciousness. To this end, I propose a test from conceptual object permanence, which can be applied not just to robots, but to the innumerable artifacts of consciousness that texture our existences.

## I. INTRODUCTION

In his 1973 paper "Robots: Machines or Artificially Created Life?" Hilary Putnam argues that there is no definitive answer as to whether robots are conscious, but that we must instead choose whether to extend this category to them.[1] This framework, at first glance, is a convincing analysis and effective response, but it fails to consider the triviality such an extension assigns to the notion of consciousness. Even if we cannot know that a robot has experiences isomorphic to those of a human, we cannot technically know this about humans, and a great deal of us choose in any case to extend consciousness to people. However, if we default to extending consciousness to anything, we cannot have certainty over a *prima facie* account. This would ultimately render the category of consciousness meaningless.

There is, however, a view that does not directly engage in the back-and-forth concerning individual minds. Consciousness, under the doctrine of extended mind theory, is considered a partially external series of phenomena. We understand ourselves and our mental processes in relation to other minds and even non-mental objects. This is no groundbreaking statement. Putnam himself proposed the doctrine of "semantic" or "natural kind" externalism in response to this dilemma, claiming that "meanings just ain't in the head!"[2]

Putnam believes that when we interact with the rest of the world, or the "natural kind," those terms are assigned meaning via interaction with the physical structures of the world around us.[3] We could not have such terms without physical inputs, and they are therefore part of the mind in some way. Elaborating on and departing from this, Andrew Clark and David Chalmers have more recently claimed that a mechanism of "active externalism" allows us to better explain how external objects function as part of the mind.[4] Via this interpretation, robots and artificial intelligence, to the extent that they are integrated into an extended consciousness, will be at the very least viewed as conscious by its other member(s). Our perception of robots as conscious does not, however, entail they are actually conscious, but perhaps that they are beings of consciousness—a physical or informational extension of human consciousness. Under this reading, consciousness cannot be something that simply is or is not.

1 Hilary Putnam, "Robots: Machines or Artificially Created Life?," *The Journal of Philosophy* 61, no. 21 (1964): 673, 10.2307/2023045.

2 Hilary Putnam, "The Meaning of 'Meaning'," *Minnesota Studies in the Philosophy of Science* 7 (1975): 144.

3 Putnam, "Meaning of 'Meaning'," 147.

4 4: Andy Clark and David Chalmers, "The Extended Mind," *Analysis* 58, no. 1 (1998): 7, 10.1093/analys/58.1.7.



Where we can see the consciousness (or lack thereof) of a robot is in the degree to which its mental existence is contingent upon the internal conscious processes of human beings and other members in a social framework. I will take the rest of this paper to explain and respond to Putnam's arguments, their intersection with Clark and Chalmers's extended mind, and put forth a small contribution of a philosophical litmus test regarding this discussion. Ultimately, I conclude that if consciousness is not internal, evidence of its existence in other minds may not be either.

I propose that we are better able to delineate the boundaries of consciousness—and therefore distinguish between the conscious and nonconscious—via a test from conceptual object permanence. This test asks whether human consciousness is necessary for the continued existence of some class of object. If the answer is yes, it can be considered an extension of said consciousness, and therefore is classified as of consciousness. I also propose that this test can be reversed, and that evidence of other human (and indeed animal) consciousnesses can be found in the offloading of cognitive burdens through external mechanical processes.

Both versions of this test can better show us the limits of our consciousness as it is represented—and as it imposes itself—in the world. If we are able to establish that our cognitive processes are extended in some manner, then we do not need access to the qualia (i.e., internal sensations) of others or direct access to definitionally inaccessible mental structures to find reason to abandon solipsism. We do not have to choose to extend the assignation of full human consciousness to robots to say that they do have conscious properties and should be treated accordingly.

## II. PUTNAM'S DILEMMA

Putnam examines various functionalist arguments for the existence of robot consciousness, including a line of argumentation by Gilbert Ryle which claims that robots can know anything a human can and therefore participate in knowledge-discussions of a substantive sort, revealing evidence of consciousness. To quote Putnam, "If knowing that *p* is having a 'multi-tracked disposition' to appropriate sayings and question-answerings and behavings…then a robot can know anything a person can."[5]

Another argument given in defense (however qualified) of artificial consciousness is the functionalist argument that "it is part of the 'logic' of psychological theories that (physically) different structures may obey

5 Putnam, "Robots," 673.

(or be 'models' of) the same psychological theory."[6] In short, something that obeys psychological principles has, at least in theory, the functional components to constitute consciousness in a human.

Offering pushback to this perhaps overly simplistic isomorphic argument, Putnam gives the example of a robot's perception of the color red. He offers the argument that "the connection between my visual sensation of red and my utterance 'it looks as if there is something red in front of me' (or whatever) is not merely a causal one."[7] Rather, it is a quale, or perceived property, of the classical variety, a purely internal phenomenon that is not necessarily sensual. This is followed by another argument that qualia have certain intrinsic properties; we could program a robot to have the opposite response to a "sensation," whereas this is basically impossible for a human—a burning stove just hurts.[8] This is, via Putnam's view, an effective argument that does not only deny the ability of humans to establish robotic consciousness but equally so the consciousness of other people through functionalism.

He concludes that this back-and-forth is largely counterproductive, and since a quale is by definition entirely internalized, it cannot be the subject of independent or objective analysis. Yet if consciousness is entirely internalized, what method do we have to know that other human beings have consciousness, much less robots? Generally for something to determine itself to be in a certain state it needs some sort of reference, one which we have for humans—namely ourselves. We do not possess however, this point of reference for a robotic consciousness. Putnam summarizes this dichotomy by stating that "the decision, at bottom, is this: Do I treat ROBOTS as fellow members of my linguistic community, or as machines?"[9] Considering that this seems to be a question of degree in some sense, it is not surprising that the solution rests in our own choosing.

To say that we should choose to extend consciousness to an automaton is, at first glance, not an unreasonable response to this back-and-forth. However, it has somewhat confusing consequences. It makes assignation virtually impossible. If we choose to say that presumably non-sentient robots are conscious, we do not leave any room for the nuance necessary to keep this position from becoming absurd.

This pitfall in Putnam's line of argumentation stems not from the structure of the dilemma he paints, but in the conclusions he draws from it. After establishing the choice we must make, he says that we

6 Putnam, "Robots," 675.
7 Putnam, "Robots," 672.
8 Putnam, "Robots," 672.
9 Putnam, "Robots," 690.



should choose to extend consciousness to robots when in doubt, as to avoid discrimination based upon "the 'softness' or 'hardness' of the body parts of a synthetic 'organism.'"[10] This argument of discrimination makes the category of the conscious very broad—too broad, I argue, to be meaningful.

For example, if we can choose to extend consciousness to a relatively sophisticated robot, what is to stop us from assigning some degree of internally generated consciousness to a simple machine or even a sculpted piece of stone if the determinant criterion is subjective. After all, both of those are involved in human consciousness, even if they are merely artifacts of it. A direct product of consciousness seems to bear its imprint. If we follow Putnam's logic that we should *prima facie* grant consciousness to robots, we should probably also grant consciousness to any machine that can manipulate its environment. If a fork or pulley, two clearly inanimate objects, can be considered conscious, what do we truly mean by conscious?

For Putnam—who would later hold that the external structures have bearing upon the internal functioning of mental processes—as the world and human technology change, so too can the psyche. It is therefore possible to say that Putnam's answer may have been satisfactory at the time, but the advent of new mind-extending (or perhaps supplanting) technology renders the answer of personal choice obsolete. However, even on Putnam's own terms, in his own time, this conclusion is ambiguous and has seemingly untenable implications. His argument would strongly benefit from some means of clarification.

All of this is not to say Putnam is incorrect in his conclusion, but the methodology he uses to establish it is open to relatively unsophisticated critique and in need of further development. Luckily, it is not necessarily the case that consciousness is entirely internalized, and the work of Putnam, Clark, and Chalmers taken together helps us to outline a more clearly exteriorized model of the mind.

## III. THE EXTENDED MIND: AN ANSWER?

To illustrate the structured fluidity that allows a notion of mental extension to provide a suitable answer, Clark and Chalmers give an example wherein three subjects are asked to rotate a shape on a screen.[11] One is asked to perform it mentally and visualize it. One may choose whether to do it themselves or press a button to have it turned on the screen. Another may do it themselves or have a robotic brain implant do it for them. There is not any fundamental difference in the way that

10 Putnam, "Robots," 691.
11 Clark and Chalmers, "Extended Mind," 7.

one would use the button or the implant, and this is given as evidence that the physical barrier of the brain cannot be equivalent with mental constructs, leaving room for the possibility of mental externality.

Furthermore, a process of coupling is laid out in which objects, particularly ones designed to aid mental processes, become inextricably linked to basic aspects of consciousness. They lay out their position fully here, writing, "Our thesis is that this sort of coupled process counts equally well as a cognitive process, whether or not it is wholly in the head."[12] Indeed, examples of this abound in our everyday lives, from eyeglasses to emails. Instead of one's qualia informing the choice to communicate or not, to respond to external stimuli or not, we can offload fairly elementary mental processes, demonstrating external coupling.

Eyeglasses are a clear example of this. Sight is one of the basic senses through which consciousness receives the inputs that allow it to maintain a homeostatic character. Without glasses a severely visually impaired person would suffer greatly—certainly they would be unable to perform many basic functions of daily life. This object, two pieces of glass and a metal frame, are included as part of the mental processes of perception at a very basic structural level.

They also give the example of Scrabble, a game in which the letter tiles being rearranged cannot be arranged in the precise physical state necessary to complete a word without including the physical tiles as part of one's thought processor—perhaps, as part of the thought. "One can," they elaborate, "explain my choice of words in Scrabble, for example, as the outcome of an extended cognitive process involving the rearrangement of tiles on my tray."[13] While they admit this could be chalked up to a series of inputs and outcomes in a mental Turing machine, they respond that "if an isomorphic process were going on in the head, we would feel no urge to characterize it in this cumbersome way. In a very real sense, the re-arrangement of tiles on the tray is not part of action; it is part of thought."[14]

However, does this exterior model hold up to the objections I have brought towards Putnam's claims? In some ways it does not. Saying that an object can be part of the mind does not represent a break from Putnam's position—it essentially is his position. While this may be true, claiming that a cognitive process can be external does represent a radical break in the discussion. However, under Clark and Chalmers's reading, consciousness could be viewed as vaguely defined unless we impose criteria that can help us sort out what is part of an extended mind's

12 Clark and Chalmers, "Extended Mind," 9.
13 Clark and Chalmers, "Extended Mind," 9.
14 Clark and Chalmers, "Extended Mind," 10.



consciousness, and to what degree we should consider that system's components to be conscious or part of the process of consciousness—functionally or otherwise.

Clark and Chalmers, therefore, do provide some sort of framework through which we can interpret different consciousnesses within a diffuse system. This is made clear in their concepts of the socially extended consciousness and even self. These are categories with some inbuilt definitionality. There are boundaries to the self that, even if we seek to extend beyond our own flesh and blood, prevent us from assigning anything to it. A rock or old-growth forest, for example, has existed and will exist independently of me. Furthermore, my broader cognitive processing and sense of integrity are not impinged on by them. But how exactly do we determine what external objects can and cannot be part of a cognitive process?

## IV. CONCEPTUAL OBJECT PERMANENCE

This leads me to a main issue around which much of this paper has been building: how do we determine what is and is not part of the extended mind, and how can that help alleviate the objections I raised to Putnam? I propose the use of a test based upon "conceptual object permanence," a phrase that needs unpacking. Object permanence is, simply put, the ability to recognize the continued existence of objects even if you cannot verify their immediate presence (i.e., sense them). This is an important development in early childhood and is often used as a test of intelligence in humans and animals.

Conceptual object permanence applies more broadly—concepts being the means of this test, not the subject. It is best to illustrate with an example. A Fitbit is something I know exists. I know it will continue to exist if I leave it in my bedside drawer or lend it to a friend (assuming they are not overly clumsy). However, if humans were to disappear, would my Fitbit continue to retain its purpose? Could Fitbits, as a kind, continue to hold meaning? One can conclude they would not, and this means they are at the very least a physical extension of consciousness.

We can also flip this test and ask about things we consider to be parts of an extended consciousness disappearing. If the multiple species of crops we have genetically modified to suit our agricultural needs were to disappear, could humans as a kind continue to exist? The answer, in this case, is that we probably could, but not very many of us. Human civilization most certainly would collapse, and we would likely revert to hunter-gatherer status, leaving behind very little that would be recognizable to you or me. This demonstrates the integral nature of something we would consider unconscious to most humans' processes

of consciousness. It seems that if a class of objects or processes are both impossible without consciousness, and that our current collective state of consciousness is impossible without said class, they are extensions of consciousness.

And what of the statues I claimed to show Putnam's position's absurdity? They do not pass this test. If all statues disappeared, this would not disrupt the functioning of anybody's consciousness as a whole. Human civilization as it currently exists would probably go on just about the same. While statues are an artifact or even arguably an extension of human consciousness, they are not integral to its functioning in a way that would make them a fundamental part of an extended cognitive process. This is the line Clark and Chalmers draw, between active and passive externality; we can see that this test distinguishes these categories accurately.

These examples may seem strange because they do not involve what appear to be immediate cognitive processes, at least not conscious ones. A Fitbit is not something I consciously manipulate through concentrated cognitive effort like a Scrabble set, and the amount of times the average person manipulates cereal crops with their mind a day is probably very close to zero. Why do these function as part of cognitive processes, then? Simply put, they alleviate the burden of consciousness. What were formerly natural cognitive processes have been offloaded onto artificial solutions. These object classes, both strictly and biologically mechanical, could not exist without us, nor we without them—indicating a process of cognitively directed mutual dependence.

## V. THE OBJECTIONS OF A SEMANTIC EXTERNALIST

Putnam might respond to this extension of his ideas by rejecting it on the grounds of the nonlinguistic nature of the test proposed. In his conclusions on the dichotomy he presents over robotic consciousness, Putnam proposes that "the decision, at bottom, is this: Do I treat ROBOTS as fellow members of my linguistic community, or as machines?"[15] This extension of linguistic credence to other minds, specifically robots, is obviously one that would seem to exclude the possibility of the extended models of cognition which I have examined and proposed.

Indeed, it may not make sense per Putnam's logic to treat any nonlinguistic actor as a fellow consciousness, or strictly of consciousness in any way. For example, clearly a Fitbit is not indicative of any original, internal consciousness in the manner that he indicated. This can even be said for his isometric robot—although he clearly does not seem to agree on this point. Pointing to the biological essentiality of consciousness (as

15 Putnam, "Robots," 690.



many philosophers, linguists, and scientists might), we could say that even if a robot can be treated as part of a linguistic community or an extended mind network, this is not evidence of some idealized, purely human form of consciousness per se.

These are both fair objections, yet they miss the distinction I seek to make. When we talk about a distributed mind and the artifacts of consciousness, we eliminate the need to talk about the fundamentally decentralized process of consciousness in strictly binary terms. One could respond to these objections by claiming that linguistic thought is simply one aspect of conscious life—there are plenty of demonstrably conscious organisms that exist without it—and that what we might want to ask is whether an aspect of an extended consciousness could reasonably be thought to have its central nucleus in some linguistic form.

However, this opens another debate that cannot be resolved within the frame of this discussion. Upon accepting an extended cognitive model, one could simply consider Putnam's linguistic demand through the lens that many examples of mind extension that do not appear linguistic are, in fact, formulated in the mind as such. This can be seen in the example of a Fitbit—we do not necessarily formulate our hunger or tiredness in linguistic, informational terms. However, the Fitbit always does this, substituting purely phenomenological experiences for explicitly textualized, and therefore linguistic, data. This could even be termed a "linguistic takeover" of thought, in which previously unarticulated, purely phenomenological, experiences are perceived in the mind in a propositional and formalized manner. Clearly, whether we can speak with some aspect of mind cannot be the only criterion for its consideration as such.

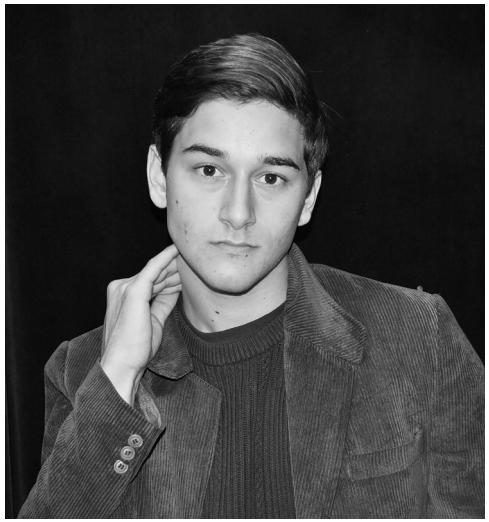
## VI. IMPLICATIONS

What implications does the theory of extended mind have upon the discussion of robotic consciousness outside of the inclusion of a robot in one's consciousness? An obvious answer is not particularly forthcoming at first. A robotic consciousness outside of our own extended mental framework is not something humans are prepared to intuit. Extended mind theory provides a middle ground upon which further research and application can be based.

Furthermore, the test of conceptual object permanence proposed here may be of some interest to those seeking an answer to the choice Putnam has laid out. While by no means complete, the general outline of a litmus test for such a model of consciousness may be useful to further research in envisioning the boundaries such a category demands.

Most consequentially, I have shown that functionalist models of the mind that focus only on the biologically internal components of human consciousness should seek to readjust upon acceptance of Clark and Chalmers's theory. Functionalists might aim to focus not just on brain states but upon how those brain states are distributed along behavioral and technological axes. If we accept that human consciousness shapes and is profoundly shaped by the world around it, perhaps the best evidence of consciousness is not something intrinsically inaccessible, but has been right in front of us this whole time.





Jacob Berk will receive a Bachelor of Arts in honors philosophy with a minor in cognitive science from McGill University, located in Montréal, Québec, in the spring of 2023. His research focuses on the intersection of and dialogue between continental philosophy and the sciences. When not waxing philosophical, he writes songs.