McGill University

Final Essay

Robots: Machines or Artificially Created Life?

Hilary Putnam

Presented by

Kaloyan Kanev - 261180884

Presented to

Professor D. Davies

In the context of

PHIL 306

Philosophy of Mind

Fall 2024

December 16thth, 2024

"*So God created man in his own image, in the image of God he created him; male and female he created them.*" If we were to believe, then should we also now believe that we have taken this as an example for our own artifice? Are our creations in our own image, are they a reflection of ourselves? Is all this cracked pottery, all these eroding walls, and all the bloated, overheating batteries simply images of our own flaws as creator of these things? We humans seem to have sought to make creations of increasing grandeur, from the Roman aqueducts that permitted the propagation of life over long distances, to the conception of our offspring in tubes of transparent melted sand, to the reproduction of exact biological copies of Finn-Dorset sheep. These past years, we have seemingly been obsessed with the idea not to reach the skies with our towers, but instead reproduce – or perhaps, create – what is intrinsic to our kind using a handful of on and off switches. The very thing that defines us as human beings – that, for many, distinguishes us from other creatures – we seek to craft using variations in current and silicon. As the topic of artificial *intelligence* becomes increasingly relevant every passing day of this second decade of our 21st century, we must wonder about what we mean by designating it as such. Why is it that we call it intelligent: is it a simple description of the machine's capacities based on human comparison; or is it out of hubris, pride, or perhaps hope, to have created beings with the most elusive and defining property of man: conscience?

Hilary Putnam, an American philosopher who extensively discussed the mind, language, and science, and who was very critical of both others' works and his own, also wondered on the topic of machines some 80 years before today, and whether we should ascribe consciousness to them; whether we should speak of them in such fashion. In his paper "*Robots: Machines or Artificially Created Life*", he essentially asks the question "Are robots conscious?" and provides arguments for why the answer to this question isn't one we can deduce or discover, but rather one that we must decide to assert. In this essay, I will thoroughly analyze his paper by looking at his various arguments, first going over how he establishes his thought experiment with the robot and its properties and capacities, then how he responds to one obvious and two relevant objections to this thought experiment. In a second part, I will analyze

how Putnam compares humans and robots on a psychological level, where he will wonder more directly whether robots are conscious. He presents the arguments philosophers give *against* robots being conscious, followed by arguments *for* robots being conscious. Finally, Putnam shows that both sides of the debate have arguments that don't hold much weight and asserts that whether robots have a consciousness or not is a matter of decision to include them in our linguistic and conceptual framework of consciousness, rather than discovery about robots being or not being conscious.

Putnam first begins by setting the scene for what will become his thought experiment. He tells the tale of robots who were many, forming a community. These robots, like earlier stages of humanity, did not know how they were made and what they were made of. Furthermore, these robots speak a language with all the linguistic characteristics that a language like English entails. Putnam ascribes to these robots the capacity to have internal states, and he describes that when something evokes the appropriate internal state for a robot – in this case seeing red – the robot would call it 'red'. For now, this seems coherent with our own capacities as human beings. We also seem to call certain things red, because these things evoke the appropriate internal state where we believe we see red. We humans also seem to know that what I see as red might not be red for someone else. We also know that if we were to be in circumstances where our vision is altered, we wouldn't trust things to be actually red, but we could only speculate that they are. For now, the robot does not have these more complex capacities.

Putnam then ascribes to the robot the capacity of inductive reasoning and theory construction. With that, the robot is capable of understanding that he could have been mistaken and that what he saw as red wasn't actually red. He would be able to think 'it looked red'. Additionally, if holding the appropriate information, the robot would also be able to understand that what appears red to him *could* not be red in actuality, allowing him to posit 'it looks red, but it isn't really red'. The robot, like us, would know that his senses are capable of deceiving him, and that there is a distinction between reality, and his perception of reality. Putnam also explains that like us, the robot wouldn't be able to speak of the

appearance of appearances, because this is simply not a logical kind of statement in natural language. 'Looks red' is what Putnam calls 'incorrigible', meaning that 'looks red' doesn't refer to the fact that we are accessing the property of red of something directly, but instead that to us, within our understanding formed by our perceptions, something *appears* to be red.

Putnam now makes a very interesting comparison between the state of robotkind in the presently described form and ancient Greeks, claiming that they would know just as little about their internal makeup, but would – just like mankind – be able to develop a scientific understanding of the world that is increasingly complex, and arrive at similar philosophical challenges as we do today. (In a tangent, I find it very amusing that Putnam states "why should a robot not be able to do considerably better than many of our students?". This statement seems very relevant in our time, where indeed many robots now seem to be doing the homework of many philosophy students. To say that they are as good as philosophy students, perhaps not yet; however, we can definitely say that they are as good as the students that request their homework be done by robots!). Putnam then shows that the robots could also arrive at the Mind-Body problem, where they too would be able to posit that "[…] 'being in the state of seeming to see something red' and 'having flip-flop 72 on' are two attributes and not one." He claims that they would be able to arrive at such a problem if they satisfy the conditions that they use language and construct theories, don't initially know their own physical make-up, have senses and can perform experiments, and come to know their make up through empirical investigation and theory construction. The goal of Putnam here is to show us that these robots, although initially starting with very little, would actually not be or come to be too different from humans now. With some very basic reflective properties that we ascribe to them, robots would be able to – just like us – arrive at philosophical problems, be forced to develop possible solutions from what they are capable of sensing about the world, and would write papers and teach classes and have students and assign final essays. Additionally, just like humans, robots would be able to bike and get hit by cars and feel pain, both localized and global. Robots – like humans – would understand that some forms of damage to their physical body could also impact their reflective capacities,

and they would develop an understanding that there must be a connection between body and mind, even though when we think we are in pain, it isn't our thoughts that are in pain. Up until now, Putnam's reasoning seems coherent, and to my limited understanding as a previously mentioned philosophy student, I would agree with Putnam's explanation, just like a robot with a limited understanding would also be able to agree with Putnam's explanation. My limited understanding further proves Putnam's point because such limited understanding is comparable to that of the described robot.

Human beings are assumed to learn to utter 'something looks red' when having the sense-perception of seeing red through observation and analysis of when other human beings utter the same thing. Additionally, we would expect human beings that correctly utter 'something looks red' in the correct circumstances to correct fellow humans when they utter 'something looks red' when something looks blue. The meaning of the utterance 'something looks red' is a "function of the rules that govern its construction and use". 'Something looks red' and 'something in my mind has gone off as a result of my visual sensing of a relatively low frequency wavelength' are two utterances governed by different rules. In correctly-behaving human beings, the statement 'something looks red' can be uttered when something in the human's mind has gone off, whether the human being is aware of that something going off or not. So when Putnam constructs this thought experiment, he is obviously surprised when one attempts to object by stating that 'the only thing this thought experiment proves is that robots are capable of only *simulating* human *behaviour*', when in fact, just like for humans, in this thought experiment, a robot utters 'something looks red' "whenever flip-flop 72 is on, *whether the robot "knows" that flip-flop 72 is on or not*". In the same way as humans, utterances are rule-governed, and the robot isn't necessarily able to give an explanation as to why he uttered what he did (in other words, doesn't know that what caused him to make the utterance was flip-flop 72 being on), but that if he were aware of flip-flop 72 being on ("from empirically established theory together with such observation statements as its conditioning may prompt it to utter, or as it may hear other robots utter"), he would be able to provide an explanation as to why he would utter 'flip-flop 72 is on'.

The other two objections that Putnam focuses on are from Baier. The first one being that humans utter because they *know* that they are having a sensation, not that their sensation merely *evokes* an utterance. Robots, instead, utter because they are *caused* to utter it by flip-flop 72 being on.

However, this thought experiment doesn't imply that whenever the robot sees red it will automatically and, as Putnam puts it, uncontrollably utter 'something looks red'. Human beings seem to be able to have some restraint and not utter X whenever they are in a state where it would be relevant to utter X. Similarly, flip-flop 72 being on should not directly cause the utterance 'something looks red'. Furthermore, Putnam correctly argues that regular (to avoid saying 'normal' people) don't talk about knowing that they have a sensation. For regular people, knowing they have a sensation of red just means they have the sensation of red. To visualize this better, we could simply think about all the times we are experiencing a certain sensation, like that of pain after a collision with a car. We are experiencing the pain, we can talk of it in the sense that we can state that we are in pain (that we have the sensation of being in pain), and this would be equivalent to 'knowing' that we are in pain. It would seem bizarre to ask ourselves, or even be asked by someone else, if we knew whether we are in pain or not. In a similar fashion, robots could talk about knowing they have a sensation of red, simply because robots can talk about having the sensation of red.

Putnam argues that the way we would talk about robots having the sensation of red differs from the way we talk about humans having the sensation of red simply because we talk about robots differently than how we talk about humans. When we talk about robots, we imply certain things in our speech about them that we do not imply when speaking of fellow humans. So, Putnam pushes the analogy further and proposes that robots do not think of themselves as robots, and that instead they believe that they have souls, just as humans have once believed that. When the robot talks about his fellow robots having the sensation of red, he would talk in the same way as we humans would talk about another human having the sensation of red: not implying that they are in mere special physical states. Just like human beings, the robots could "fail to find 'correlates' at the physical level for the various sensations they report".

Putnam pushes the analogy *even* further, hypothesizing that the robots could develop ROBOTS of their own, and the robots would talk about their ROBOTS in the same way that we humans talk about robots. The robots would, in the same way as Baier, argue that their ROBOTS are in special physical states when this ROBOT would think that something is red or that something looks red to a ROBOT.

The second objection is that qualia have intrinsic sensation properties, and that a quale Q that we find unpleasant cannot be pleasurable to another human experiencing the exact same quale. Robots, however, can be reprogrammed to feel pleasurable when experiencing a qualia Q that was initially unpleasurable to them. Baier claims that for this reason, the robots' internal states don't have intrinsic sensation properties to them.

Putnam quickly replies to this objection by showing that "if [the physical] correlate [of some pleasurable psychological state] is a highly structured state of the whole brain, then such reprogramming may well be impossible". Also, it relies on the assumption that qualia are intrinsically pleasant or painful.

In his next part, Putnam will assume that humans and robots are psychological isomorphic, in the sense that they have a "sameness of functional organization". Putnam clarifies that when he says that robots have a psychology (that they obey psychological laws), he doesn't imply that robots are necessarily conscious.

Putnam wonders about Oscar, a robot among many. When Oscar has the sensation of red, Putnam wonders if Oscar sees, if he thinks, feels; if he is alive, if he is conscious. After a lengthy establishment, we are brought back to the core question. He refers to this problem as the "civil rights of robots", because he claims that we would get to such a topic of discussion much faster than we would expect. Once again, it is very interesting to be reading about this in our time, in a time where robots (in the sense, machines, or computer programs executed on machines) are becoming increasingly complex, performing increasingly human-like tasks, and are expected – or, at least, assumed – to very soon reach a status of self identification. To Putnam, this seemed close, but was still very far away, because at the time of him

writing this paper, technology was nowhere near what it is today. Today, this problem is not a problem yet, but might become one tomorrow. (It is also important to note that I don't mean *problem* in the normative, negative meaning of the term).

In this part, Putnam looks at various arguments *against* robot consciousness, followed by arguments *for* robot consciousness, stating that he is more interested in the way we speak of robots; the 'semantical aspects of our language'. He will show that both sides of the debate provide arguments that are without merit.

The first anti-civil-libertarian argument Putnam exposes as a bad argument is the phonograph-record argument, which states that a robot only "plays" behaviour in the sense in which a phonograph record plays music. When we are told a joke by a robot, if we were to laugh, we would laugh at the wit of the human programmer who programmed the good joke in the robot, we aren't laughing at the robot itself. This argument assumes that the robot cannot learn, and that it is simply programmed with a limited and predetermined set of behaviour routines. This is not only uninteresting, but also isn't psychologically isomorphic to the human, as we had previously stated we would assume. If instead the programmer had programmed the robot so that it will be a model of certain psychological laws, we would have trouble predicting the robot's behaviour, as much as we would have trouble predicting a human being's behaviour in real-life situations, even if we knew every single psychological law perfectly. On jokes, if we were to correct the argument so it follows the pre-established psychological isomorphism, the robot would have been programmed to *produce* new jokes, not a robot that has been programmed with a predetermined human-written set of jokes.

Another objection that I would personally give to this argument – even before Putnam's response that this argument assumes no psychological isomorphism – is that at times, perhaps because of a certain amount of programmed realism, I tend to suspend my disbelief and laugh at not only the joke, but the fact that it is the machine that tells me the joke. Sometimes, I don't find the joke as amusing as the very idea that this joke was told to me by a machine that I probably assumed to be of no good to my

amusement. In cases like that one, I don't object that it is a robot that told me the joke. I would probably simply object at the joke being very funny. But I believe that with enough programming or a bit more effort in making a really good joke, I would probably be able to reach levels of amusement I find myself experiencing with human beings. Additionally, even though this would imply further programming, my suspension of disbelief would still be in action, and what proves that I would still believe that I am appreciating the joke coming from the robot is my surprise at the fact that such a good joke I would hear from a machine. I wouldn't have such surprise and be in such awe if I knew and asserted that "Anyways, the joke was programmed, and the wit is that of the programmer".

The second argument against robot consciousness is the reprogramming argument, which states that a robot has no real character of its own, for it could at any time be reprogrammed to behave in a different way. In contrast, a human that would be reprogrammed through a form of brain operation, which would give him a new and completely predetermined set of responses, would no longer be a human being in the full sense, and would instead be a monster.

Putnam's response to this is to give an example of a criminal that we would "reprogram" through some brain operation for him to become a good citizen, without destroying the criminal's capacity to learn, develop, change (potentially also change back to a criminal), then we would certainly not have created a monster. If we were to assume psychological isomorphism (which we do), robots would be reprogrammable to the same extent as humans would be.

I would also like to argue that the reprogrammability argument seems dubious to me simply because it therefore assumes by contradiction that humans aren't reprogrammable. But this is very far from the truth if we look at simple examples such as psychological therapy. A more complex example would be to look at the neuroplasticity of the brain, which allows us (through intensive speech pathology) to reintroduce some linguistic capabilities which were initially lost through brain damage (Aphasia) in specific areas of the brain which were originally responsible for these same linguistic capabilities (namely Broca's and Wernicke's area). This is reprogramming in its most literal meaning.

The third argument against robot consciousness is the question-begging argument, stating that what we call psychological states of robots are just physical states, but that ours aren't, so robots aren't conscious. This argument assumes that psychological predicates don't apply to robots and to humans in the same sense, "which is just the point at issue".

Putnam states that all these arguments against robot consciousness suffer from the same flaws, namely that they assume robots are artifacts and that they are deterministic systems of a physical kind, whose behaviour is preselected and designed by an artificer, assuming that these properties are *not* properties of human beings. In other words, it is as defendable to claim that humans aren't conscious as it is to claim that robots aren't. If it is illogical in the English language to speak of robots as conscious beings, if saying 'Oscar is in pain' or 'seeing a rose' or 'thinking about Vienna' is a violation of a rule in English, we have not been told what rule this violates.

Putnam then confronts the arguments *for* consciousness of robots. It is argued by some that terms like 'anger' or 'sensation of red' are implicitly defined by psychological theories. However, if these terms depend on theoretical constructs, their truth relies on the correctness of the theories underlying them. but since these theories aren't empirical, they are contingent, and could therefore be false, without altering the meaning of the terms. Psychological terms are argued to have a reporting use, like that of reporting that 'I am in pain'. For a psychological theory to be valid, it would require empirical consequences, but these aren't linked to ordinary psychological terms. For example, linking pain to observed behaviour such as a cry doesn't fully explain pain as a concept across different contexts. So, similarly to the previous arguments against Oscar's consciousness, if it is illogical to speak of Oscar as not being conscious, in other words if saying Oscar is not a conscious being violates a rule in the English language, we have also not been told what rule is being violated.

Putnam then analyzes Ziff's actually relevant argument against robot consciousness. Ziff claims that there is a semantical connection between being alive and being conscious, so if we were to prove that Oscar is not alive, we would also prove that Oscar is not conscious. Ziff argues that when we say

that something is alive, we don't do it by describing that something's behaviour, because the movement of the hands on my clock don't make it alive, or the growing of a flower in my garden doesn't make it alive if it turns out to be made of gears and springs that set it in motion. In this case, it is (finally) a violation of semantical rules of our language to say that something is alive if it is clearly a mechanism. Ziff therefore argues that it is structure, not behaviour, that determines whether or not something is alive.

Putnam replies to this objection by stating that there is a lack in nuance in Ziff's argument. In fact, he argues, that we could create an anthropomorphic machine, made out of flesh-like, "soft" materials. Think of "Blade Runner"-esque replicants. In this case, we would still have a robot, but this time it wouldn't be a mere mechanism made of gears and bells and whistles. This becomes paradoxical, because on one hand we have a being that we assume to be non-conscious on the basis of its mechanical structure, but on the other hand, we have a non-conscious being that is made out of human-like materials. In other words, two things with identical psychological behaviour would be judged differently based on their material composition, in one case unconscious because hard, and on the other, conscious because soft. Ziff argues that the English language does not permit us to speak of a thing made of mechanical parts to be conscious, just as it is impossible to speak of a married bachelor. However, it isn't as unintuitive as Ziff claims it is, because we are definitely capable of imagining a conscious robot. In fact, we have been imagining and writing about them for decades. So, there isn't anything about the rules of the English language that wouldn't permit us to talk about a conscious robot. Putnam therefore asserts that Ziff is wrong, and that although it might be false, it is not a contradiction to say that Oscar is alive.

Finally, Putnam discusses the Know-Nothing view, which maintains that it is just as possible for robots to be conscious as it is for them to not be. The arguments can take a theological or non-theological form. "My body happens to consist of flesh and blood, but it might just as well have been a machine, had God chosen." Humans are composed of body and soul. A soul could animate a mechanical body such as the one that Oscar possesses. However, since Oscar – and robots in general – seem to be animated without a soul, it is just as plausible that Oscar – and robots in general – are lifeless, soulless machines. The non-

theological form of this argument essentially replaces the soul with the mind. To understand a statement like "Oscar is conscious", we must first know what it is like to be conscious ourselves. However, understanding the statements does not guarantee they have truth values. The problem that arises from this is that we cannot claim that humans are conscious, even if they appear to be similar psychologically or physically to ourselves. We cannot argue using analogy because we could also imagine robot philosophers using the same arguments from analogy to talk about each other and talk of their own ROBOTS as we talk of our robots. Being able to imagine Oscar as conscious does not prove that statements like 'Oscar is conscious' have a truth value.

For these reasons, Putnam argues that Oscar's consciousness isn't something that we have to – or even can – discover, but it is rather about making a decision about how to treat Oscar – and robots in general. If we make the choice to treat robots as part of our linguistic community, concepts like 'consciousness' and 'alive' can then be meaningfully applied to them. If we cannot, such statements remain meaningless.

Putnam concludes with the rather tragic realization that we fail to provide the relevant proof of a consciousness – or lack thereof – in robots, which would allow us – or bar us – to speak of such a property when speaking of machines. And since we cannot prove it, because we have failed to discover it, we can only resort to *deciding*, whether to ascribe consciousness to machines, or not. It becomes not a matter of scientific objective understanding, but a choice we have to make of essentially discriminating what is and what is not conscious. And since we can only base our discriminations on explainable physical properties, it becomes just as undefendable as other physically based discriminations, like racism.

Perhaps the way we choose to talk about machines, and whether we would choose to ascribe to them a conscience, an intelligence, or not, doesn't say much more about them, but says more about us. "*What are today only philosophical prejudices of a traditional anthropocentric and mentalistic kind would all too likely develop into conservative political attitudes.*"