Can a robot with artificial intelligence have free will?

This chapter discusses the question of whether there could ever be artificial intelligence with free will. This question reduces to ~~the question of~~ whether or not artificial intelligence will ever be able to act on domain-general rationally formed intentions. However, it might not be possible to distinguish between behavior that is produced by such intentions and behavior that is merely a simulation. This could be because, as the Chinese Room thought experiment seems to show, consciousness intuitively is a necessary condition for intentionality. There are three potential responses to this challenge~~worry~~. One could argue that full domain-general rational behavior is possible only with consciousness, or that it might also be necessary for the system to be implemented in the right substrate, or that the right microfunctionalist structure needs to be in place for real intentionality.

artificial intelligence, domain-general, intentions, consciousness, simulation, Chinese Room, microfunctionalism

Chapter 11

Can a robot with artificial intelligence have free will?

Jonathan Hall and Tillmann Vierkant

To answer the question of whether an entity such as a robot with artificial intelligence (AI) can have free will, we will lean on our chapter where we asserted that to have free will is to have *the capacity to form and act in line with intentions*.[1] An entity with this capacity we will call an intentional agent, allowing us to reframe the question of the title as "Can a robot with AI or even the disembodied AI in a computer be an intentional agent?"

[1] See Chapter 7 by Hall & Vierkant on degrees of free will.

To be an intentional agent is to interact with the environment in a goal-directed way, consistent with the agent's representational and motivational attitudes to certain propositional content. By logically combining these intentional states, the agent forms goals that drive behavior. For example, an agent may rationally form the intention to open the fridge if she believes it contains beer and she (really) desires a beer.[2]

If the capacity to rationally combine intentional states to form intentions is necessary and sufficient for free will, then determining whether robots have intentional states would help answer our question, ~~but to rule out confusion, one clarification is in order???.~~ Presumably, we can already imagine a scenario where in certain narrowly defined domains the actions of a robot require intentional explanations, but in all known cases so far robotic abilities across a wide range of domains remain far below human standards. Clarifying that our aforementioned *capacity to form and act in line with intentions* must be domain-general on a human scale will get us the result that robots even with AI are unlikely to be considered free agents in the next couple of decades, but beyond that it would be brave to rule out that robots could ever meet that criterion.

Assuming that one day the attribution of domain-general intentionality is ~~both~~ necessary to explain their behavior ~~and domain-general~~, would that be sufficient to conclude that robots have free will? Intuitions might begin to diverge here, but for many the answer might still be no. One justification for the persistence of the skeptical intuition is that attributing intentionality on the basis of externally observable actions leaves open the possibility of confusing a simulation of agency with the real thing.[3] This mirrors a debate sparked by Turing (1950) in his paper "Computing Machinery and Intelligence."

---

[2] See Chapter 1 by Yaffe on intention and Chapter 15 by Sinnott-Armstrong on reasons.

[3] See Chapter 10 by Bayne on behavioral experiments.

Turing proposed that instead of asking the question “Can machines think?,” one should consider whether there are imaginable digital computers which would do well in what he described as the “imitation game.” In this game an interrogator asks questions to both a machine and a human and, from the answers, tries to determine which is which. John Searle (1980) famously argued through his Chinese Room thought experiment that the kind of proficiency, in converting inputs to outputs, shown by an entity passing the Turing test does not imply intentionality. In particular, Searle argued that if he was in a room receiving questions in Chinese and diligently followed English instructions regarding how to respond with Chinese characters, then he could appear to an interrogator to be conducting an intelligent conversation even though “I don’t speak a word of Chinese.”

Although there has been much debate about these thought experiments, those in the Searle camp insist that one cannot infer intentionality from behavioral output. A response is needed if one is ever going to be justified in attributing agency to robots. There are a number of forms that this response could take.

First, Searle is very clear that some machines can think, because, after all, “we are precisely such machines.” Intentionality, in his opinion, must therefore be substrate-dependent, and it happens to be the case that humans are made of the right stuff. It then becomes an empirical question as to whether other substrates can support intentionality. Theoretically, an alien could have or ~~indeed~~ a robot could be built with a chemical and physical structure that supported intentionality (although it is not clear how we would know). Here we must differentiate between a moving and sensing robot with AI and the disembodied AI in a computer. In Searle’s opinion, substrate-independent computational operations on formally specified elements are never sufficient for intentionality.

Second, a powerful response to Searle is that he is making an unjustified step from the fact that certain parts of a system don’t understand Chinese to the assertion that the system as

a whole does not. Even if Searle doesn't understand a word of Chinese, it is not obvious that the combined "Searle+instructions" system does not. No one would argue that individual neurons in the brain understand this essay, but hopefully the neural system as a whole does. Although, in our opinion, this objection is powerful enough to allow that robots with AI could be intentional agents, it is unlikely to persuade the skeptic.

Also plausible is a third approach, which Clark (1991) calls "microfunctionalism." This approach agrees that formal operations at the coarse-grained symbolic level of the Chinese Room experiment can't produce intentionality, but argues that formal relations at a fine-grained sub-symbolic level could provide the right kind of structure to support flexible domain-general behavior and the attendant emergent properties associated with intentionality. In this model, the cognitive system is assumed to be "at root a sub-symbolic system" that is scaffolded by intentional states at a higher level. If this is correct, then it is not the substrate that matters but the sub-symbolic system. This provides a framework in which it could be legitimate to claim that a robot with AI is an intentional agent.

There is not enough space in this essay to adjudicate among these three positions, but our personal hunch is that it seems much more likely that intentionality and free will are to be found in the organization of the substrate rather than the substrate itself, which makes options 2 and 3 the most likely contenders. Both of these are consistent with the possibility that robots with AI have free will.

Follow-Up Questions

Mengmi Zhang

What are the relations between intentionality and free will? Do they refer to the same thing? In theory of mind in philosophy, mental states like beliefs, desires, and perceptual experiences have intentionality in the sense that they represent or are about things or states of affairs. Does free will refer to a specific case of intentionality?

Deniz Arıtürk

Would robots with AI ever pass your requirement of "no external interference?"[4] If so, how? What, if anything, distinguishes the external interference on the actions of robots with AI (namely, that they are built by humans) from external interferences on the actions of humans (such as that their genes are a product of natural selection)?

Antonio Ivano Triggiani and Mark Hallett

Modern technologies show that the creation of an artificial brain is possible. Still, it's hard to reach a high degree of connections similar to that among human neurons, and it's also difficult to emulate their plasticity. So, does free will in a machine depend on its complexity?

Gabriel Kreiman

I really enjoyed reading this lucid answer. My question is, basically, what sort of *empirical data* would lead us to think that a robot has or does not have free will?

Let us start with behavior and consider a Turing test for free will. In room A, there is either a machine or a human; in room B there is either a machine or a human. You can ask *any* question. Based on the answers, how would you determine which room has an agent with free will?

I suspect that you will not like this formulation of the question. You may argue that behavior is not enough; there has to be "intentionality." I would like to make sure that the definitions connect to *empirically observable variables*. If behavior is out (or insufficient), then I will allow you now to record the activity of every single neuron in a biological agent and the voltage of every transistor in the robot. Feel free to add whatever variables you think are relevant here—calcium concentration in the pre-synaptic terminal, the position and composition of every atom. If there is an empirically measurable variable that you want, you

[4] See Chapter 7 by Hall & Vierkant on free will in degrees.

got it! Based on those measurements, how would you determine which room has an agent with free will?

If the answer to these questions is that there is no empirically observable variable that would ever determine which agent has free will, then "intentionality," "free will," and similar terms have little scientific value as we cannot falsify them, we cannot measure them, we cannot use one in any empirical way to assess the other, etc. Then it seems that we have made up a specialized vocabulary with no connection to the physical world. For example, I could argue that certain flies have wtx3xtw. If I build a robot fly, would it have wtx3xtw? Never! Because it would lack zy6yz! Can we measure zy6yz! No. Can we measure wtx3xtw? No. But I shall assert that flies with zy6yz certainly have wtx3xtw.

David Silverstein and Hans Liljenström

Can a deterministic system, like (presumably) a computer or a robot, have intentions or free will? Suppose actions from free will are based on freely determined intentions. If algorithms for intentions are modeled deterministically, how can goal-directed decisions from environmental inputs be based on free will? Will these not be pre-determined? Some aspects of human experience may appear random and may partially drive the development of intentions. If modeled intentions have a random component, can that help represent free will, or does this just dilute it? To what extent are intentions driven by self-models?

Replies to Follow-Up Questions

Jonathan Hall and Tillmann Vierkant

A number of questions were concerned with the different meanings of "intentional" (e.g., Zhang; Silverstein & Liljenström). We are happy to clarify these. Intentions as executive states that drive behavior are a subgroup of the wider category of intentional states that also encompasses, for example, beliefs (often referred to as Brentano-intentionality or aboutness).

As we say in the text, in order to form intentions in the former sense, it is necessary that an agent can rationally combine intentional states in the latter sense (including beliefs). The biggest obstacle for allowing that robots with AI might have free will is that they do not seem to have any intentional states in the latter, Brentano sense.

An additional problem is that even if robots behaved as if they had Brentano-style intentionality, there is a big debate on whether that means they really do have it. In a way, the same problem exists obviously for humans, but in our own case we have intuitive first-personal access, and we assume that other humans with very similar brains to our own will probably also have it. We do not want to judge how important this evidence is, but it explains the intuitive difference between robots and humans.

Given this clarification, it is now much easier to answer the other questions. In a way, we here have reduced the question of whether robots can have free will to the question of whether robots could have intentional states in the Brentano sense that they could rationally manipulate to form intentions. There is a large literature in the human case on whether the having of intentional states is compatible with having a designer (Arıtürk), or what kind or degree of complexity is required (Triggiani & Hallett), or whether there could be an empirically observable variable to test for the existence of this (Kreiman). All we can do here is say that the answers to these questions in the human case will allow us to answer the robot case as well.

The one specific claim that we can already deduce from this is that randomness is not a major factor for free will on this account (Silverstein & Liljenström) because it seems unlikely that randomness will play a major role in explaining intentionality in humans. This outcome is hardly surprising, though, because we started off with a compatibilist notion of free will in the first place. So our answer does not address whether robots could have incompatibilist free will, but it does ask interesting questions of the incompatibilist. Would a

robot that possesses aboutness and the ability to form intentions in a rational way really be not free if randomness played no role in its decision-making? And would it change anything if we added a randomness generator to the robot that could influence its decision-making in some way?