Res Cogitans

Volume 6 | Issue 1 Article 7

5-29-2015

Artificial Intelligence, Zygotes, and Free Will

Katelyn Hallman
University of North Florida

Recommended Citation
Hallman, Katelyn (2015) "Artificial Intelligence, Zygotes, and Free Will," *Res Cogitans*: Vol. 6: Iss. 1, Article 7. http:dx.doi.org/10.7710/2155-4838.1124

Artificial Intelligence, Zygotes, and Free Will

Katelyn Hallman
University of North Florida

Published online: May 29 2015

Abstract

In this paper, I assume that strong AI is possible and I question whether AI robots would have free will. The ultimate goal of this paper is to use our intuitions regarding AI and free will to motivate incompatibilism. I argue that AI cannot act freely because the nature of an AI robot's design keeps it from being able to have the kind of control required for free will. The strategy of this paper is to first define the control condition of free will. Then I discuss Mele's Zygote Argument and compare it to AI. Then I briefly discuss advancements in AI technology and briefly describe how AI would work. Next, I show how an AI machine cannot satisfy the requirements for free will. Following this, I use these intuitions to motivate incompatibilism—the concept that free will is not compatible with determinism. Finally, I respond to a series of objections. It is my hope that, using this AI thought experiment, we can come to a similar conclusion that Mele came to: AI is not relevantly different from humans, AI would not have free will, so neither would humans (*if* humans turn out to be determined in the appropriate way).

As we move into the age of technology, one goal that humanity has been striving for is the creation of artificial intelligence (AI). AI, according to the Association for the Advancement of Artificial Intelligence, is "the scientific understanding of the mechanisms underlying thought and intelligent behavior and their embodiment in machines."[1] This is known as 'weak AI:' AI that is not sentient and has limited functionality. Strong AI, on the other hand, is AI that would be sentient, would not have limited intelligence and functionality, and would be able to do all of the things that humans could do. In this paper, I assume that strong AI is possible and from that starting point I question whether AI robots would have free will. Specifically, could something that is programmed[2] have free will? The ultimate goal of this paper is to use

[1] "AI Overview: Broaddiscussions of Artificial Intelligence." *Aitopics.org.* Accessed December 1, 2014. http://aitopics.org/topic/ai-overview

[2] We can consider programming to be the both theological and causal determination. Programming would be theological determination because it requires an intentional creator with

our intuitions regarding AI and free will to motivate incompatibilism. I argue that AI cannot act freely because the nature of an AI robot's design keeps it from being able to have the kind of control required for free will *and* that humans are not unlike AI in a way that would give them free will.[3] The strategy of this paper is to first define the control condition of free will. Then I discuss Mele's Zygote Argument and compare it to AI. Then I briefly discuss advancements in AI technology and briefly describe how AI would work. Next, I show how an AI machine cannot satisfy the requirements for free will. Following this, I use these intuitions to motivate incompatibilism—the concept that free will is not compatible with determinism. And finally, I respond to a series of objections.

I. What does it take to have free will?

In this section, I define and motivate three definitions of the control condition of free will. I discuss free will as choosing on the basis of one's desires, free will as having alterative possibilities, and agent causation.

1. S freely does A iff
   i. S acts on basis of S's desires and values.
   ii. S's desires mentioned in (i) are desired desires and values.

According to this definition, a person can have free will if they do something because they want to do something. For example, a person giving an armed robber their very expensive and special family heirloom despite their wishes would not be acting freely in this instance.[4] But, additionally, the desire must also be desired. Therefore, a person who has a drug addiction that wishes they did not have the desire for the drug does not freely take drugs. This condition excludes those whose desires are subverted by something else.

2. S freely does A $=_{Df}$ S could have done something different.

a plan setting up the actions of the being. Programming would be causal determination because the program and laws of computing would be analogous to the laws of nature and other causal factors.

[3] I'm not committing to the idea that determinism is true for humans. But we all should agree that it is true for robots. I merely argue that if determinism is true then, using my AI thought experiment, humans would not have free will.

[4] One could argue that the person was doing what they wanted in that they wanted to not be killed, and the desire not to be killed outweighed their desire to keep their family heirloom. I would not disagree with that. However, we cannot ignore the fact that the person still does not want to give away their possessions regardless of what other desires may outweigh this desire.

So in the case where one person decides whether to make a decision, in order for that person to have free will they must be able to have done or chosen something different. There are multiple ways of interpreting this requirement for free will. One interpretation is that if they had desired to do something different, they could have done it. Reconsider the example of the person giving the armed robber their family heirloom. If they were to act freely, they would've had to want differently and to do something different based on that want. They could've had the desire to fight back, and fought back because of that desire—that action, according to this definition, would've been done freely. Another interpretation of this requirement is having the ability to do something and not being restrained from doing it. Consider a situation where you are brainwashed and your mind is put under the control of some evil scientist, and this evil scientist controls everything you make. In this situation, you would be restrained—you would be unable to freely move about the world and make your own decisions and movements—you would be unable to act freely.

3. S freely does A iff
   i. S's action does not originate from an outside source – it originates from S.

This definition of free will requires that the agent—the person exerting the free will—be the source of the action in question. If we were to draw a causal chain, mapping backwards from the event to the source, the chain would have to end with the agent as the source according to this view. The agent must make the decision to act, independent of other events, followed by a brain event, muscular movement, and so on.

Another interpretation of this requirement for doing an action freely is that while there may be factors that make one more likely to decide one thing over another, one can still act freely. Those who support this view make the following claim: the link between previous factors and a decision could merely be probabilistic (in which case the factors wouldn't wholly determine the decision) and this would still allow for an action to be done freely and simply because the decision *belonged* to the agent.[5]

Of course, there are other ways of spelling out the requirements for free will, but I feel that any other definition of free will could fit into one of these three categories. Consider another definition of free will: S freely does A iff S's is unconstrained and is not inhibited in an unusual way. But this could be drawn back to agent causation and limiting one's possible alternatives. For example, if an agent were hypnotized to act a certain way, the actions from the agent henceforth would originate from the one who hypnotized them—not the agent. Being constrained does not originate from the agent itself and limits ones possibilities of action.

[5] O'Connor, Timothy & Edward N. Zalta (ed.) "Free Will", *The Stanford Encyclopedia of Philosophy* (Fall 2014). http://plato.stanford.edu/archives/fall2014/entries/freewill/

II. Mele's Zygote Argument and AI

We could consider AI to be an extreme version of Mele's Zygote Argument. In his thought experiment, Mele hypothesizes the creation of a zygote (the result of two sexual reproduction cells combining), named Ernie, in which the genetic material is manipulated in such a way that 30 years later Ernie will "will judge, on the basis of rational deliberation, that it is best to [do] A and will [do] A on the basis of that judgment, thereby bringing about [event] E."[6] Every other action Ernie takes, however, will not be determined/manipulated/controlled. From this, Mele questions whether this manipulation precludes Ernie of free will only in that action or in every action. I claim that we should all agree that in the very least that Ernie was not free in performing A. Mele, however, takes it further and concludes that this manipulation precludes Ernie of all free action based on the claim that the manipulation to do A (which causes event E) is no different than the normal 'blind forces' of nature that go into the formation of a zygote.[7]

I argue that AI is similar to Ernie in the zygote argument—but, unlike Ernie, every action of AI is directly manipulated/controlled, rather than just one. In this paper, I hope to show that AI machines do not act freely at all. I will show that the fact that AI is not biological doesn't matter, and it not being intentionally programmed doesn't matter. It is my hope that, using this AI thought experiment, we can come to a similar conclusion that Mele came to: AI is not relevantly different from humans, AI would not have free will, so neither would humans (if humans turn out to be determined in the appropriate way).

III. AI: What it is And How it Works

The first technology considered 'artificial intelligence' were programs that could solve mathematical computations, programs that could play games, and industrial robots. Now, artificial intelligence has advanced to include intelligent personal assistants (e.g. Siri, Cortana, Google Now, etc.), facial and image recognition software, self-driving cars, etc. The AI technology we have now mainly focuses on a few aspects of intelligence and would be considered weak AI. Image recognition software only focuses on perception and interpretation of images, and intelligent personal assistants focus on understanding language and data mining, etc.

These advancements, while huge, are only part of the goal of attaining artificial intelligence. There's more to humanity than just conversation and image recognition. A strong AI robot would be able to think and act in the same way that humans think and

[6] Mele, Alfred R. "Manipulation, compatibilism, and moral responsibility." *The Journal of Ethics* 12, no. 3-4 (2008): 279.
[7] Ibid., 280.

act. We would be unable to distinguish their thoughts and behavior from human thoughts and behavior. AI would have to be able to perceive and distinguish objects, describe scenes, pick out objects in scenes, etc. AI would have to be able to reason using both logic and emotion. Emotion, personality, tastes or preferences, memory, etc. would be an integral part of AI's capabilities.

So, how would AI do these things? Well, we ought to remember that an AI robot would essentially be a computer and so it would operate in the same way as a computer. The 'brain' of every computer is called the Central Processing Unit (CPU), and the CPU is split into two parts: the Control Unit and the Arithmetic/Logic Unit (ALU). The control unit sends the directions from the program to other parts of the computer so the program can be carried out and the ALU performs all mathematical and logical operations that the program might require. The CPU also has access to the memory storage—the location where the program information is stored. Memory is broken down into two categories: primary and secondary. Primary memory is memory that is permanently stored in the computer whereas secondary memory is memory that is only temporarily stored in the computer and is deleted once the program or computer is shut down.[8]

Computers can only do what it is stored in the program—if it is not in the program the computer cannot do it. Programs give computers step-by-step instructions that are executed one line at a time. A CPU with only one control unit can only execute one instruction at a time, although they can execute instructions extremely quickly.[9] However, there are three ways you can speed up a computer or enable computers to execute multiple instructions at a time: add additional CPUs, hyper-threading, or add more control units or "cores" to the CPU. An AI machine would have to be *extremely* complex in order to emulate personality, decision-making, unconscious thoughts and desires, etc.

IV. AI and Free Will

In this section, I detail why AI cannot have free will. I draw connections between the facts from sections I and III to show that AI does not satisfy the provided definitions of free will.

Recall the first definition of free will: An AI robot has free will iff the AI robot acts on basis of its desires and values and the AI robot's desires are desired desires and values. This definition of free will requires that the robot act based on its desires and values.

[8] Fay-Wolfe, Vic. "How Computers Work: Disks And Secondary Storage." *http://homepage.cs.uri.edu/*. Accessed December 1, 2014. http://homepage.cs.uri.edu/faculty/wolfe/book/Readings/Reading04.htm
[9] Ibid.

An AI robot could certainly have desires and values—it could be programmed to prefer certain sights, sounds, people, political ideologies, etc. A programmer could even add the extra layer of making the desires and values desired. An important aspect of humanity is the ability to self-reflect and the ability to reflect on one's beliefs and values—which is a requirement for the second condition of this definition. Since an AI robot would have to be able to reflect on itself and its beliefs, it should be able to have desired desires. But, consider the case where the robot has undesired desires—how would it change the desires? Given that a computer cannot do anything other than what it is programmed to do, a robot could only change its desires if it were programmed to change those desires. Additionally, the self-reflection and evaluation of the desires would have to be programmed, so the robot would be programmed to desire some desires and not desire some desires. Everything, even when it comes to desires and desired desires, is traced back to the program.[10] While those who hold this view of free will would not feel let down about this fact, I argue they ought to. They would say that so long as the robot does what it wants, it is free. This may sound intuitive and well, but I argue that we must also consider where the desires come from—the robot's desires are technically not its own because they can be traced directly to the programmer. Additionally, the AI robot had no other possible alternatives for desires; once the desires were programmed, the robot has no other alternative (unless it of course is reprogrammed). It seems odd to say that free actions can be caused by something that was not freely gotten (i.e. something that we have no control over).

Recall the second definition of free will: An AI robot has free will $=_{Df}$ the AI robot could have done something different. As we discussed in the previous section, a computer cannot do something that it is not programmed to do. Anything an AI robot does will already be programmed—given the way the program was written, the AI robot could not have done anything differently. Assuming that it is even possible to program an AI robot to be able to respond to the large unknown (perhaps infinite) number of possible situations it may encounter, the robot would have to somehow make a decision. There are two ways an AI robot could make decisions: random selection out of a preset list of alternatives or choosing from a preset list of alternatives based on its desires.

Consider the robot having a list of preprogrammed options to choose from when confronted with a possibility for action and randomly selecting which option to take. While this may sound alright, randomness is not synonymous with free will. Randomness seems to imply that you are not the one making the decision—if nothing

[10] If we were to translate this into the language of the free will/determinism debate, I would say that every desire would be traced back to the laws of nature and the complete physical state of the universe. The case of AI and free will is not very different from the general free will/determinism debate.

you want and if nothing in your past guides your decisions, then that's the same thing as letting a dice roll rule your life.[11]

So, if randomness that doesn't work, perhaps if the robot had wanted to do something different it would have done differently. But this doesn't work either, because anything an AI robot would want to do would be programmed. An AI robot couldn't want something that it wasn't programed to want—so it couldn't really do anything other than what it was programmed to want to do. But again, those who hold this view would not feel let down about this fact, though I argue they ought to. They would say that as long as the robot could have done something different *if* it wanted to do something different, it is free. But I argue this is inadequate for the same reason in the first requirement rejection. We must also consider where the desires come from—the robot's desires are technically not its own because they can be traced directly to the programmer. The robot may have been able to do differently if it had wanted differently, but its desires again don't really belong to the robot—so it seems unfair to grand it free will even in this case.

Recall the third definition of free will: An AI robot has free will iff the AI robot's action does not originate from an outside source—it originates from the robot itself. But a computer cannot do what it is not programmed to do. Every act a program does can be found in the program, and someone writes every program; therefore, any action taken by an AI machine can be traced back to a source outside the program—everything a program does not only *can* be directly traced back to the programmer, but it *needs* to be directly traced back to the programmer. A programmer is necessary for having a program because if we don't have a programmer, we don't have a program.[12] Additionally, *everything* an AI robot would do, think, believe, want, etc. would be found in the program, which was necessarily created by a source outside of the robot.

But, there being a programmer does not guarantee that there is a program; you could have a programmer who is very lazy or not motivated and therefore never writes a program, etc. So, in order to preclude it from having acted freely, we must find something outside the robot that provides a sufficient condition for the robot's actions. In order to do this, we must first find a sufficient condition for the robot's actions. The program, necessarily written by an outside source, guarantees which actions the robot would take. While programs can be written in different ways to achieve the same end,[13] the robot still has to follow its program exactly. Therefore, the program being written a

[11] Lloyd, Seth. "A Turing test for free will." *Philosophical Transactions of the Royal Society: A Mathematical, Physical, and Engineering Sciences*, 370, (2012): 3599-3560.
[12] An AI machine would require **non-biological** (hence the usage of the word "artificial" in AI) hardware and software to be arranged in a very specific, complex, and intentional way.
[13] There are multiple routes one could take to program the same action a computer takes (e.g. for-loops and while-loops).

certain way is sufficient for the robots actions. Now since the source of the program is outside of the robot, the source of the action must also be outside of the robot as well. The programmer (outside the robotic agent) set up the robots actions, desires, decisions, etc. (via. program) in a way such that the agents' action is guaranteed. Therefore, we can conclude that is a sufficient cause of the action (the program) that originated outside the agent.

V. Using AI to Motivate Incompatibilism

Incompatibilism, one response to the problem of free will, is the claim that that one could *not* have free will in a deterministic system—that free will is incompatible with determinism. Incompatibilists do not claim that determinism is true; rather, incompatibilists claim that *if* determinism were true, then we would have free will.

I argue that AI cannot be considered to have 'acted freely' if everything it does was decided/determined by something other than itself. Specifically, I define a completely deterministic system one in which the conjunction of all of the laws of nature and factors in the past provide sufficient conditions for (i.e. guarantee) the actions/events at time t—as is the case with AI. As I have showed in the previous sections, program necessitates the actions that every computer takes and a computer cannot do any action that it was not programmed to do. Every AI machine's desires, personality traits, possible choices, etc. would have to be explicitly defined in the program—a program that was necessarily written by a programmer (an outside source). AI would fail each control condition for having free will: choosing on the basis of one's desires, having alterative possibilities, and agent causation. It seems then, that if we follow these intuitions, we ought to conclude that AI would not have free will. I consider these intuitions regarding AI and free will as support for incompatibilism. AI would not have free will because of the necessary and sufficient causal link between programmer, program, and action. Furthermore, the programmer and the program are analogous to the laws of nature and other factors relevant to determinism; therefore, if determinism were true then humans would not have free will either.

VI. Objections

AI, though very similar, would still be different from humans – how does AI not having free will preclude humans from having free will? Specifically, AI would be artificially (not-biologically) created, and AI would have an intentional creator that created the machine and the laws the govern it (it's an open ended question as to whether humans do). Could these factors preclude AI from having free will while still allowing humans to have free will?

What does being artificially created have to do with free will? Well—nothing, really. Something that is artificial is simply something not naturally occurring or created in a process that does not occur naturally. If we don't allow AI to have free will based on the fact that they are created in an artificial way, we will have to exclude those conceived by in vitro fertilization or artificial insemination. The more difficult question is whether AI being non-biological excludes it from the free will category. The problem of free will is concerned with the question of whether something (whether biological or not) could have free will even if everything was causally determined. In this paper, I showed how something that is clearly causally determined and very similar to humans would not have free will—and AI not being biological should have no impact on that conclusion. Something being biological does not change whether it is causally determined.

Furthermore, AI having an intentional creator is also irrelevant. It would only be relevant in the case where having an intentional creator was the *only* reason AI wouldn't have free will. But AI would not have free will because of the fact that every AI machine's desires, personality traits, possible choices, etc. would have to be explicitly defined in the program—a program that was necessarily (and sufficiently) created by an outside source. Without the program (and programmer), AI would not be able to act (necessity condition); the program (and programmer) guarantees certain AI actions (sufficiency condition). This is what precludes AI from having free will, not having an intentional creator. Humans wouldn't have if they lived in a deterministic system—one in which the conjunction of all of the laws of nature and factors in the past are both necessary and sufficient conditions for the actions/events at time t—regardless of whether they have an intentional creator.

VII. Conclusion

AI, since it would essentially be a *very* complex computer, could *not* do anything that is not in its program and would not be able to have free will. Since AI cannot control its program or the fact that its program *entails* all of its future actions, personality, beliefs, etc., it can also have no control over anything it does and thus does not act freely. It would be, in essence, similar to an AI chess partner. While AI would have some very sophisticated decision making programs, intricate personality programs, self-reflexivity, etc., it will still, at its core, be programmed. Its beliefs, desires, desired beliefs, desired desires, every possible choice it might make, will all be found in its source code—in the program that was programmed by a necessary and sufficient source outside of the robot (the programmer).

Reconsidering Mele's Zygote Argument we can see how AI would be an extreme version of this thought experiment. Mele compares Ernie's manipulation to the natural 'blind forces' that occur when a zygote is formed, and claims that the scientists

manipulation is no different from the blind forces.[14] From this, Mele concludes that Ernie would never act freely. Similarly, AI involves the same type of manipulation that Ernie had, except with AI the programmer would manipulate every action and decision the AI robot would take. Additionally, as I discussed in my objections section, it does not appear that there are any relevant differences between AI and humanity that would preclude AI to have free will but still allow humans to have free will. Therefore, I conclude that incompatibilism is true but leave it an open question as to whether humans are causally determined in the same way that AI would have to be.

References

Fay-Wolfe, Vic. “How Computers Work: Disks And Secondary Storage.” *http://homepage.cs.uri.edu/*. Accessed December 1, 2014.

http://homepage.cs.uri.edu/faculty/wolfe/book/Readings/Reading04.htm
Griffin, Andrew. “Turing Test breakthrough as super-computer becomes first to convince us it's human.” *The Independent.* June 8, 2014.

http://www.independent.co.uk/life-style/gadgets-and-tech/computer-becomes-first-to-pass-turing-test-in-artificial-intelligence-milestone-but-academics-warn-of-dangerous-future-9508370.html

Hoffman, Chris. “CPU Basics: Multiple CPUs, Cores, and Hyper-Threading Explained.” *How-To Geek.* August 8, 2014. http://www.howtogeek.com/194756/cpu-basics-multiple-cpus-cores-and-hyper-threading-explained/

Johansson, Linda. "The Pragmatic Robotic Agent." *Techne: Research In Philosophy & Technology* 17, no. 3 (Fall 2013): 295-315. http://dx.doi.org/10.5840/techne2014249

Lloyd, Seth. "A Turing test for free will." *Philosophical Transactions of the Royal Society: A Mathematical, Physical, and Engineering Sciences*, 370, (2012): 3579-3610. http://dx.doi.org/10.1098/rsta.2011.0331

Mele, Alfred R. "Manipulation, compatibilism, and moral responsibility." *The Journal of Ethics* 12, no. 3-4 (2008): 263-286. http://dx.doi.org/10.1007/s10892-008-9035-x

O'Connor, Timothy & Edward N. Zalta (ed.) "Free Will", *The Stanford Encyclopedia of Philosophy* (Fall 2014). http://plato.stanford.edu/archives/fall2014/entries/freewill/

[14] Ibid., 280.

"AI Overview: Broaddiscussions of Artificial Intelligence." *Aitopics.org.* Accessed December 1, 2014. http://aitopics.org/topic/ai-overview