The Nonexistent Moral Agency of Robots – A Lack of Intentionality and Free Will

1. Introduction

In this paper I will address the concerns mentioned in Sullins' article "When is a Robot a Moral Agent?" In doing so, I will support the argument that robots are not, and cannot be, moral agents; robots lack the intentionality and free will necessary for moral agency because they can only make morally charged decisions and actions based off of what they were programmed to do. First, I will provide my interpretation of the contentious notions of "intentionality," "free will," and "moral agent." Second, I will present an exposition of arguments relevant to this paper. The arguments discussed will be those presented in John P. Sullins' "When Is a Robot a Moral Agent" and Selmer Bringsjord's "Ethical Robots: The Future Can Heed Us," in which they support and oppose, respectively, the concept that robots can be moral agents. Third, I will provide a critical analysis of Sullins' argument for the moral agency of robots. This will be done through exposing the ambiguities and faulty reasoning present in his argument. Fourth, in response to Sullins' comments on Bringsjord's argument, I will argue that humans are not "programmed" the same way as robots are. I will further clarify that the intentions of humans are not exclusively influenced by their environment through the use of findings from a twin study conducted in the field of cognitive psychology. This will ultimately support the argument that robots are not moral agents, in which the desires and intentions of humans sets them apart from robots. Finally**,** I will review the purpose of this paper and the conclusion reached.

2. Exposition

2.1 Intentionality, Free Will and Moral Agents

In this section, I will provide an explanation of how the concepts of "free will," "intentionality" and "moral agent" will be utilized in the context of this essay, as intentionality and free will are a necessary condition of being a moral agent. With this, I will further consider competing views while providing reasoning against them.

The use of the term "intentionality" in this paper will be based off of the explanation developed by John Searle; "intentionality is that property of many mental states and events by which they are directed at or about or of objects and states of affairs in the world" (Searle 1). Specifically, intentionality only encompasses certain mental states, such as fears, desires, hopes, beliefs or other states that are "about something" (Searle 1-4). In other words, mental states of intentionality are those that *attend to* objects (Jacob). These intentional states experienced by individuals are results of, and fulfilled in, structures in the brain, as "Intentional states stand in *causal* relations to the neurophysiological… and Intentional states are *realized in* the neurophysiology of the brain" (Searle 15). Therefore, mental states of intentionality are caused by, and produce other, biological phenomena (Searle 264). With intentional states explained, it can be further concluded that an intentional action is a way by which an individual sets out to accomplish, or the "conditions of satisfaction" of, an intention (Searle 80).

Free will and intentionality work together, as the existence of intentionality proves that an individual has free will (O'Connor). As stated by Johnson, "intending to act is the locus of freedom; it explains how two agents with the same desires and beliefs may behave differently" (Johnson 200). With this, I will now move to explain the notion of free will in light of how it will be used for the remainder of this paper. Taking a general approach to the notion of free will, it can be assumed that "free will has two aspects: the freedom to do otherwise and the power of

self-determination" (O'Connor). As this notion is merely surface level and lacks specificity, the "freedom to do otherwise" is more concretely described as the possibility that, if one had hoped for something different, they would have pursued that end instead (O'Connor). Additionally, the "power of self-determination" is assumed in the case that "an agent self-determines her ϕ-ing just in case ϕ is caused by her *strongest desires* or *preferences* at the time of action" (O'Connor). Thus, free will, generally speaking, encompasses one's freedom to do otherwise and ability to realize they acted out of intentionality.

In accepting that intentionality, and therefore free will, exist among certain individuals, it can be further entailed that such individuals are moral agents. I will utilize the line of thought developed by the "causal theory of action" when referring to moral agents, as specific conditions must be met in order to demonstrate moral agency (Schlosser). First, the agent of an action must have mental, internal states, those of which include the agent's intentions (Johnson 198). Second, the agent must act externally, and such external action is caused by the mental, internal state of the agent him or herself (Schlosser). Third, the action caused by the agent has an impact on the external world (Schlosser). Finally, this act specifically "[harms or helps]" another individual, who is considered to be the "patient" of the situation (Johnson 198). In light of this, the readings referred to and points made in this paper will primarily focus on the first condition of moral agency, in which a moral agent's actions are caused and defined by their mental state. Thus, whenever an action carried out by an agent is explained through "referring to their beliefs, desires and other intentional states," they have the potential of being considered a moral agent (Johnson 198).

2.2 Arguments Surrounding Robotic Moral Agency - Bringsjord versus Sullins

With explanations of important terms as a foundation, I will now establish the opposing arguments provided in Bringsjord and Sullins' articles to present the issues to be discussed. I will start by unpacking Bringsjord's argument, in which he provides reasons suggesting that robots are not moral agents. In his article, "Ethical Robots: The Future Can Heed Us," Bringsjord makes the greater argument that "we ought to fear not robots, but what some of us may do with robots" (Bringsjord 539). Within this argument, Bringsjord establishes that a robot, such as the one in his lab, will only perform actions based on its coding (Bringsjord 542). The particular robot he works with, PERI, can perform both moral and immoral actions by, respectively, performing actions of holding onto or dropping a ball representing Earth. PERI is unable to choose such action on the basis of its morality, as individuals controlling it are ultimately the ones who make those decisions (Bringsjord 542).

In addition to this, Bringsjord carried out a second experiment to further demonstrate that robotic free will, when "actions performed correspond to those that are provably advisable, where 'provable' is fleshed out with help from standard deduction over knowledge represented in the situation calculus" is not possible (Bringsjord 542). Such robots as PERI will follow instructions of the "prover," which is ultimately a set of rules (Bringsjord 542). The prover is built by humans, therefore meaning PERI is *still* run by humans and does not have robotic free will. This includes the fact that even if a "random factor" were added to the program to make the robot do something "surprising" to the programmers, its actions are still "determined by some random factor, not freely chosen by the machine" (Sullins 156). Finally, Bringsjord clarifies that through being objects ultimately programmed by humans, robots will always be controlled by

humans. Thus, through this, Bringsjord concludes that robots are not moral agents and do not have autonomous will, because they are restricted to what they are programmed to do.

In response to Bringsjord's argument, Sullins claims that "[humans] are all products of socializing," and he continues to elaborate on the concept that humans are thereby "programmed" by their environment (Sullins 156). The environment experienced by humans is composed of a system of behaviors which is transferred from individuals non-genetically. Sullins then goes on to state that that "if Bringsjord is correct, then we are not moral agents either, because our beliefs, goals, and desires are not strictly autonomous" (Sullins 156). He elaborates on the fact that humans beliefs, goals, and desires are a result of various elements of our environment, including culture and education (Sullins 156). Thus, Sullins comes to the conclusion that our environment is a form of "programming" and that humans are not moral agents in light of Bringsjords account, thereby demonstrating an apparent flaw in Bringsjord's argument.

Following his argument contra-Bringsjord, Sullins outlines the three necessary conditions for a robot to be considered a moral agent: autonomy, intentionality and responsibility. In order for a robot to be autonomous, Sullins states that it must not be under "direct control" of an individual. To further clarify this, Sullins differentiates between practical independent agency and effective autonomy. The former merely has the level of a telerobot, which is not considered autonomous. Robots with effective autonomy hold a higher level of autonomy, as such robots pursue their goals, thereby allowing them to have potential for moral agency. Robots are thus specifically moral agents when they have effective autonomy and their agency "causes harm or good in a moral sense" (Sullins 158).

Yet autonomy is not sufficient on its own as a condition of moral agency. Sullins adds that a robot must act with intentionality if it is to be considered a moral agent. This means that the robot must be acting in such a way where its behavior is "complex" with regard to leaning on its predisposition or intention "to do good or harm" (Sullins 158). With this, Sullins remarks that the robot is a moral agent "if the complex interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial and the actions are seemingly deliberate and calculated" (Sullins 158).

Finally, Sullins presents responsibility as a necessary condition to be present alongside a robots autonomy and intentionality. He assumes the robot to be responsible in situations where the "robot behaves in such a way that we can only make sense of that behavior by assuming it has a responsibility to some other moral agent(s)" (Sullins 159). In order to be a moral agent in the condition of responsibility, the robot must be in a *social position* that holds responsibility, as well. To clarify the condition of responsibility, Sullins states that "these beliefs, or programs, just have to be motivational in solving moral questions and conundrums faced by the machine" (Sullins 159). With this, Sullins comes to the conclusion that so long as the robot is autonomous with actions backed by its intentionality and responsibility, it must be considered a moral agent.

3. Critical Analysis

3.1 Analysis of Sullins' Requirements for Robotic Moral Agency

In order to put forth the argument that robots are not, and cannot be, moral agents, I will start by arguing against Sullins account; I will demonstrate that there exist ambiguities in his requirements of robotic moral agency and apply his argument to the self-driving car. Sullins'

argument is flawed, as he relies on the requirement of "intentionality" for robot moral agency without explaining exactly what he takes intentionality to be. In other words, he fails to legitimately outline the concept yet still chooses to use it as a requirement. In his article, Sullins provides that one is acting intentionally when "the behavior is complex enough that one is forced to rely on standard folk psychological notions of predisposition or intention to do good or harm" (Sullins 158). This definition ultimately leaves one to question which "standard folk psychological notions of predispositions" Sullins is referring to (Sullins 158). One may defend Sullins through pointing out that he *does* add that the complex behavior and "interaction of the robot's programming and environment causes the machine to act in a way that is morally harmful or beneficial and the actions are seemingly deliberate and calculated" (Sullins 158). This would allow for less ambiguity surrounding his explanation of robot "intentionality" (Sullins 158). Yet the use of the word "seemingly" causes his explanation to remain ambiguous, as one may question the grounds upon which an action is considered "seemingly deliberate and calculated" (Sullins 158). Thus, the grounds upon which he defends the intentionality of robots is ambiguous, as he fails to provide a solid framework for what one should consider to be an intentional action. This demonstrates that Sullins' requirement of intentionality cannot utilized in determining the moral agency of a robot, thereby providing that the way in which he argues for the moral agency of robots is ineffectual.

This conclusion can be further suggested through an example of the self-driving car. With respect to Sullins' three necessary conditions for a robot's moral agency, one would assume the self-driving car is autonomous machine with intentionality and responsibility. This is particularly evident in the fact that the machine is "to autonomously decide who should live and who should

die… without real-time supervision" (Awad et al. 63). Self-driving cars are autonomous in the sense that they are not remotely controlled by another individual. They have intentionality, as defined by Sullins, because driving is a way one interacts with an environment in a morally harmful or beneficial way. On a similar note, they hold a large amount of responsibility, in which they must ensure the safety of both the individuals inside the car alongside other agents on the road. Thus from the perspective of Sullins, one could conclude that the self-driving car is a moral agent.

Yet through disregarding the requirement of intentionality provided by Sullins and applying the notion of intentionality described in the exposition of this paper, it is evident that the self-driving car cannot be a moral agent. This is apparent as, Edmond Awad et al. conducted an online experiment to "explore the moral dilemma faced by autonomous vehicles" and determine the "ethical principles that should guide machine behavior" (Awad 59). The mere fact that there exists such a study to determine what would be the morally right course of action for a self-driving car prior to its creation suggests an impossibility of assuming robots can be moral agents. This is because such debates would likely not be necessary should the self-driving car possess mental states, specifically beliefs, about morally charged situations such as a car crash.

3.2 Arguing Against Sullins' Reply to Bringsjord

Now that I have established weaknesses present in Sullins argument surrounding the moral agency of robots, I will now provide further support for Bringsjord's argument. I will do so through arguing against Sullins' reply to Bringsjord. Sullins compares the programming of robots to the "programming" of humans, and he further uses this comparison to argue that

Bringsjord's argument would mean humans aren't moral agents themselves. Yet such a comparison between the programming of humans and robots should not be made.

To begin, I will look at the influence of the environment on humans and robots. The environment does interact with human genes in a superficially similar way to how it does with the programming of a robot. It has been established that "genetic instructions, in conjunction with environmental influences, produce a phenotype, an individual's physical, behavioural and psychological features" (Kail 52). Depending on the environment one is in, certain behaviors are manifested. Thus, human behavior is *in part influenced* by environmental factors; the environment indirectly causes humans to behave the way they do. In a similar light, robots are influenced by their environment, as they react to their surroundings. For example, the self-driving car "UniBwM" can detect crossroads and determine parameters of an intersection, therefore allowing its environment to determine where and when it can move (Dickmanns). Therefore, from this standpoint, one might assume that the environment equally influences humans as it does robots.

Yet, there are two primary reasons by which humans are different in their "programming" from genes and interaction with their environment, which sets them apart from robots. First, parents do not currently *choose* every "program," or gene, of their child. This therefore means that humans are, in a way, implicitly programmed to act the way they do. There does not exist and individual who *explicitly* determines the genes, and therefore ones mental states. In contrast, as established by Bringsjord, robots are explicitly programmed to act as they do, as the programmer determines all of the possible functions of a robot. Even if the robot is programmed

to include a "random factor," the choice to include this function was ultimately made by the programmers (Sullins 156).

Second, genes are generally considered by psychologists to play a large role in the "psychological phenotype" of an individual (Kail 63). This is especially apparent, because in the field of behavioral psychology it has been stated that "across the animal kingdom, individual differences in behavior are nearly always influenced by genetic factors which, in turn, result from a substantial number of individual genes, each with a small effect. Nearly all genes that affect behavior influence multiple phenotypes" (Kendler and Greenspan). From this statement, it can be concluded that the behaviors, and thereby mental states, generated by individuals are greatly determined by one's genetics. Furthermore, although one may believe that we are "products of socialization," the results of twin studies conducted in the field of behavioral psychology can be utilized to display the role of genetics despite one's environment (Sullins 156). These studies are conducted to determine whether the cause of certain behaviors can be attributed to genes or the environment. In these studies, comparisons are made between identical twins and fraternal twins, as identical twins have the exact same genes whereas only about half of the genes between fraternal twins are the same (Kail 61). When the identical twins result in being "more alike" than fraternal twins in a study, one can conclude that their behavior is hereditary (Kail 61). With this in mind, a twin study conducted by Plomin and Crabbe exemplifies how genes can play a significantly large role in the programming of an individual. It was found that identical twins "show a remarkable similarity for depression: If one identical twin is depressed, the other twin has roughly a 50 percent chance of being depressed," whereas there is only approximately a 25 percent chance of two fraternal twins to be depressed (Kail 63). The

findings of this study, and other similar studies, thereby demonstrate that genes can determine one's behavior despite their environment.

In contrast to point mentioned above, robots are completely dependent on the environment to carry out actions in the way they do. As an example, the "UniBwM" self-driving car is entirely dependent on its surroundings, in which it recognizes the "parameters of the driving lane and the two neighboring lanes" in order to stay on the road (Dickmanns). The car will not "choose" to drive outside its lane as a result of its programming (Dickmanns). This ultimately demonstrates the difference between human versus robot interactions with their environment.

In light of the notion of free will as explained in the exposition of this paper, the two points discussed above demonstrate that humans have free will and intentionality, or are moral agents in the sense that is presented by Bringsjord. As humans are implicitly programmed by their parents, they are not designed to act in a specific way. In other words, robots are explicitly programmed to act in the way they do, whereas humans possess the free will because they have the freedom to do otherwise and the power of self-determination. The fact that genes play a large role in human behavior, despite one's environment, exemplifies that one's mental states are not entirely influenced by their surroundings. This further demonstrates that individuals are not controlled by their environment in the way that robots are, and therefore humans act with intentionality.

4. Conclusion

In this paper I provided an interpretation of "intentionality," "free will," and "moral agent." From there, I provided the arguments surrounding the moral agency of robots, in which Sullins and Bringsjord presented opposing views. I further elaborated on Sullins' reply to Bringsjord to demonstrate possible reservations one may have with regards to robots *not* having moral agency. With the two specific arguments of robot moral agency presented, I then rendered Sullins argument ineffectual. This allowed for the more structured notion of intentionality to be applied to the moral agency of robots, as that of Sullins could not be utilized. Following my analysis of Sullins' requirements of moral agency, I took greater consideration of Sullins' argument contra-Bringsjord. In doing so, I presented that the human interaction with the environment and the "programming" of humans by their genes could not be equated with those of robots. This further focused on the requirement of free will and intentionality for moral agents. Overall, it was ultimately demonstrated that humans differ from robots, in that humans possess moral agency, through free will and intentionality, whereas robots do not.

Works Cited

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Bringsjord, S. (2008). Ethical robots: the future can heed us. *AI & SOCIETY*, *22*(4), 539–550. https://doi.org/10.1007/s00146-007-0090-9

Bringsjord, S., & Schimanski, B. (n.d.). What is Artificial Intelligence? Psychometric AI as an Answer, 7.

Dickmanns, Ernst D. “Developing the Sense of Vision for Autonomous Road Vehicles at Unibwm.” *Computer*, vol. 50, no. 12, 2017, doi:10.1109/MC.2017.4451214.

Jacob, Pierre, "Intentionality", *The Stanford Encyclopedia of Philosophy* (Spring 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2019/entries/intentionality/>.

Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204. https://doi.org/10.1007/s10676-006-9111-5

Kail, Robert V. *Children and their development*. Pearson Education, 2012.

Kendler, Kenneth S., and Ralph J. Greenspan. “The Nature of Genetic Influences on Behavior: Lessons From ‘Simpler’ Organisms.” *Am J Psychiatry*, 2006, p. 12.

Leisman, Gerry, et al. “Intentionality and ‘Free-Will’ from a Neurodevelopmental Perspective.” *Frontiers in Integrative Neuroscience*, vol. 6, 2012. *Crossref*, doi:10.3389/fnint.2012.00036.

O'Connor, Timothy and Franklin, Christopher, "Free Will", *The Stanford Encyclopedia of Philosophy* (Summer 2019 Edition), Edward N. Zalta (ed.), forthcoming URL = <https://plato.stanford.edu/archives/sum2019/entries/freewill/>.

Schlosser, Markus, "Agency", *The Stanford Encyclopedia of Philosophy* (Fall 2015 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2015/entries/agency/>.

Sullins, J. P. (2011). When Is a Robot a Moral Agent? In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics*(pp. 151–161). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511978036.013

Weissman, D. (2018). Autonomy and Free Will: Autonomy and Free Will. *Metaphilosophy*, *49*(5), 609–645. https://doi.org/10.1111/meta.12333