Acting vs. Being Moral: The Limits of Technological Moral Actors

Aaron M. Johnson
Electrical and Systems Engineering Department
University of Pennsylvania
Philadelphia, Pennsylvania, USA
Email: aaronjoh@seas.upenn.edu

Sidney Axinn
Philosophy Department
University of South Florida
Tampa, Florida, USA
Email: axinn@temple.edu

***Abstract*—An autonomous robot (physical or digital artificial being) may be capable of producing actions that if performed by a human could be considered moral, by mimicking its creators actions or following their programmed instructions, but it cannot be moral. Morality cannot be fully judged by any behavioral test, as the answer to moral questions is less important than the process followed in arriving at the answer (as evidenced by disagreement among ethicists on the correct answer to many such questions based on the individual's moral style). The distinction between acting and being moral was recently considered for lethal autonomous military robots [1], and in this paper is further clarified in the context of more broad applications. In addition such a distinction has implications for what types of tasks autonomous robots should not be allowed to do, based on what must be moral decisions. Here we draw a distinction between what might be illegal for an agent to do (which relates more to the agreed upon laws of the current political leadership), and what actions are so innately moral decisions that we cannot delegate them to a machine, no matter how advanced it appears.**

I. Introduction

Humans react better to robots that are anthropomorphized, with life-like appearance or behavior, which allow for more familiar and comfortable interaction [2]. Indeed designing a robot's behavior to be more human like may be an unavoidable key to the widespread adoption of such technology in human environments [3]. However when a robot acts like a human does we may be tricking ourselves into believing that it is really thinking like humans do. When it appears to have emotions it is only natural for the people it interacts with to attribute the emotions that they would have in a situation to the robot. These behaviors may be a beneficial factor in their social integration, but it is important to be clear that reproducing such actions is different from experiencing the underlying emotions. Commanding a robot to appear sad does not mean that the robot is experiencing sadness in the same way that humans do. The physical expression is a consequence, not the source, of experiencing an emotion.

Similarly the consequences of morality are often observed through actions, but those actions are not the source of the morals. Instead it is the internal decision to take a certain action. In this way morality requires free will – it is the choice to do the right thing instead of the wrong thing. This first requires the capable of doing the wrong thing, which itself raises concerns [4]. But more critically a programmed robot cannot make such a choice; instead it may only do that which its algorithms dictate. Such algorithms may be quite complicated, including e.g. neural networks, machine learning, or stochasticity, but still will result in the outcome of the code's execution applied to the sensory input. Against a given example input it might produce the same action as an independent observer would. The observer might think that the robot appeared to act morally, but that is quite different from the robot actually making the moral decision to choose right over wrong.

In this paper we start by exploring this difference between being moral and acting moral, in Sections II-A and II-B, respectively. This leads to the question, in Section II-C of if there can ever be a test of morality that a robot could hope to pass. We conclude that an autonomous robot can only mimic moral actions, and therefore in Section III-A we propose a restriction on the use of robots in situations where acting upon a moral decision is necessary. This is surely not the only ethical restriction on the use of such technology, but is the focus of the present paper. As such we have found only a short, though quite important, list of roles that the proposed restriction would apply to. It is thus apparently only a major hurtle in the adoption of autonomous systems in a few situations, which Section III-B takes to be an optimistic view on the potential value of new technology more broadly in the future.

Throughout the course of this paper we will use the term "robot" to refer to any autonomous cyber-physical system. This then excludes tele-operated systems such as remotely piloted drones, which certainly have legal and moral questions to be considered but are outside the scope of this paper. The focus here is on actions, and hence the use of the term robot (whose algorithms have the ability to sense and act upon the world) and not simply computer program, although many of the arguments apply equally to both.

We also limit the scope of this paper to conventional, pre-singularity [5] levels of intelligence where the algorithms that control the operation of the system may be quite complicated, but still execute finite programs and are formally equivalent to functional evaluation. In a hypothetical future where artificial intelligence has become so advanced that it is truly distinct from this sort of program and could be considered a person in its own right then this topic, among many others, will need to be revisited. However for now such levels of intelligence lie in the realm of science fiction.

This paper to appear in the 2014 IEEE International Symposium on Ethics in Engineering, Science, and Technology.

II. Moral Actors

A. Being Moral

First we shall consider the characteristics of a human person, as a basis for comparison. This concept of a person is taken from Immanuel Kant's work on the subject, "personality is the characteristic of a being who has rights, hence a moral quality" [6, p. 220]. For the foreseeable future, any robot in question will surely not meet this requirement and therefore not be considered a person. However even if they are not a person they might still have the moral reasoning facilities equivalent to that of a person (though we argue in the sequel that they cannot).

One aspect of a person that is particularly relevant here is the imperative of morality, and that, "the concept of freedom, which points in the direction of the concept of duty, is that of a person" [6, p. 227]. To Kant a person has duties indicated by the categorical imperative. The categorical imperative is the demand that you, "act as though the maxim of your action were, by your will, to become a universal law of nature" [6, p. 39]. This presupposes that such persons have free will, and can decide whether to follow the categorical imperative, or instead to follow a selfish principle. According to Kant, to freely choose to follow the categorical imperative is to impose on one-self the principle of morality.

Even if a person is not a Kantian they may still certainly be considered moral and simply follow a different moral style [7]. Rather than considering all of humanity to be the moral benefactor they may be utilitarian and consider some of humanity to be the moral benefactor, or consider God or country to be the ultimate benefactor. Whatever the style of morality followed, it is that the person chooses to follow it and make some sacrifice for their moral benefactor that makes it a moral action. It then appears that to freely choose to follow a moral style more generally is to impose on one-self the principle of morality.

The freedom to make this choice is an oft debated topic in philosophy as, "the possibility of freedom cannot be directly proved, but only indirectly, through the possibility of the categorical imperative of duty" [6, p. 213]. Furthermore, "the concept of freedom emerges from the categorical imperative of duty" [6, p. 227]. If one is considering whether or not to obey the demands of duty, one apparently has the free will to do either one.

Stated another way, a major feature of rational beings is that, "everything in nature works according to laws. Only a rational being has the capacity of acting according to the conception of laws, i.e. according to principles. This capacity is will" [8, p. 29]. It appears therefore that in order to act according to principles, as required by the categorical imperative, we again find the requirement for free will.

A person has rights, duties, and free will, and imposes on him or herself the categorical imperative. In addition, as the Kant scholar Jane Kneller has put it, "...for the possibility of moral motivation, the imagination is indeed strangely but obviously at the root of all human experience" [9, p. 161]. For moral motivation, as Jane Kneller has said, one must additionally have the imagination to consider the effect on a person of various different actions. This requirement of imagination goes beyond the question of free will – to satisfy the categorical imperative one must have both the facilities to consider the result should the maxim of your actions be a universal law, and the freedom to change your actions based on that result.

Kneller is far from the only Kantian to emphasize the role of the imagination. Bernard Freydberg explains that, "the moral law, its forms, and the maxims that flow from it, are one and all synthetic a priori judgments and therefore include imagination" [9, p. 120]. And Fernando Costa Mattos has emphasized the, "imagination, guided by morality" [9, p. 138]. Imagination therefore appears to be a prerequisite for morality. However imagination by its nature requires a novel consideration of the world, different from the way one has in the past. Robots may be quite good at simulating many possible laws of physical interaction, and can incorporate parametric or stochastic variations on those laws. However they cannot by their nature consider anything truly novel – beyond the classes of variations preprogrammed into their algorithms.

In summary, a person has not only rights, duties, free will, but also the imagination to understand the effect of different actions, and the ability to impose on him or herself the categorical imperative. How close do robots come to the features of a human person, the features that make for moral motivation and moral action? Such robots certainly do not have rights. They do have programmed commands that seem at first to be close to the concept of duties. However a duty is not a command that must be followed; rather it is a desirable choice that one should follow. Robots do not appear to have free will – if they did we might call them "out of control." While they may have a range of choices that they consider in a given situation, that range is specified by their programs, not by themselves. Imagination requires one to consider the world not as it is, but as it could be, while robots can only consider the classes of world models allowed by their algorithms. Without free-will and imagination, they cannot impose the categorical imperative on themselves; they cannot consciously sacrifice selfishness for morality. And, if they cannot do that, they cannot be moral.

B. Acting Moral

In a stage play good actors will convey the emotions of the character that they are portraying. By controlling their voice and movements they can tell a story about what that character is feeling and how they are experiencing the world. To do this well the director may have told them to pause before a certain line, or follow another actor with their gaze. However we would not say that they are necessarily experiencing that emotion, they are simply acting it out, especially if they were told what actions to take by the director.

If their character does something amoral, say stealing money from another character, we would not blame the actor. The action taken by the actor is the same as that of a street thief, however they were simply acting out the amoral action that the writer put into the script. The writer telling an actor to act out an amoral action does not mean that the actor isn't moral. Similarly programming a robot to follow a certain action does not mean that the robot is moral. Neither has considered whether the principle of their actions could be generalized. Neither has chosen a moral style, or considered

sacrifice for the good of themselves, their nation, their religion, or for all of humanity.

A stage actor is working from a script, which generally allows for only one set of outcomes. A robot is working from its programming, and its programmed rules of morality given by its programmer is also fundamentally limited – they cannot possibly cover all scenarios, for that would take infinite memory. The program might be quite complicated, evaluating thousands or millions of scenarios, and so it may appear to have many options to choose from. Ultimately, however, the robot can act out only those actions that have been programmed into its finite memory, and of those it will choose the action dictated by its program as a result of the given input.

Another simplified example of something that gives the appearance of a moral action is a video or other recording, "when playing back a video of a moral act, one would not say that the video was moral; it is simply replaying the moral act of its subject" [1, p. 135]. An autonomous robot is obviously much more than a recording, as it can interact with the physical world. But since the robot is programmed to respond to a certain situation in a way that appears moral, it is also obviously not choosing to follow a moral style or choosing to make a sacrifice.

C. Test of Morality

Some have argued that since humans are notoriously imperfect moral actors, a programmed system could eventually be more moral than a human system [10]. However this notion that there can be a "test" of morality that a robot could hope to pass (or score higher than a human), is inherently flawed. An action can only be considered moral, according to Kant for example, if the actor choses to follow the categorical imperative. There may be common agreement about what the result of that choice is in some circumstances, however it is the choice and not the result that can be considered moral.

Furthermore there is not always common agreement about what the correct "moral" choice is in ethical problems. A moral actor must give themselves the moral imperative by choosing some moral style [7]. Someone who is a Kantian (takes all of mankind as the moral beneficiaries) may disagree with someone who follows a religious style (who takes God, or gods, as the benefactor). Both are considered to be moral actors they are simply following a different moral style. In this way it is not the answer to a moral question that is moral or not but rather the process by which one comes up with the answer.

Morality is not simply following the law. For example the laws of war are in part a set of rules, and therefore a test on what is included in them or not is certainly possible. Just because morality requires one to follow the laws of war, acting within this set of rules does not make one's actions moral. Furthermore the laws of war are not only a set of written down rules, as stated in the U.S. Army Field Manual 27-10, "although some of the law of war has not been incorporated in any treaty or convention...this body of unwritten or customary law is firmly established" [11, p. 4]. Thus the written laws are not the entirety of what a moral actor must consider, and so even a test on the ability to follow the written laws of war would be incomplete.

III. Limits of Technological Use

A. Disallowed Uses of Autonomy

In light of robots inability to be moral actors, they should therefore be excluded from performing certain roles.

1) Soldiers: The clearest actions that robots should not be allowed to do is kill people. As argued in e.g. [1], the decision to take a human life is inherently a moral decision. Fortunately there are few places where humans routinely take such actions, first and foremost being in war. Certainly in war humans have always used technology to assist in killing each other, however it has always been the human making the decision and the human who initiates the action. As technology progressed from swords to crossbows, guns, missiles, and now drones the physical distance between the attacker and the attacked has grown. But while the missile for example is self propelled, its propulsion is launched by a human, and its target trajectory selected by a human.

While the battlefield is a different environment from general life, there are no fewer variations in the infinite possible situations a robot soldier might find themselves in. Even if their use was restricted to a certain task [10], say building clearing, the possible scenarios it could encounter are still innumerable and therefore nonprogrammable. The humans that program and launch the robot cannot decide a priori how the robot should handle all of these scenarios or even all of the possible actions the robot might need to take. As it cannot make a moral judgment it must not be given the power to decide to attack a human. The use of autonomous robots in war must be limited to actions which do not require moral decisions, such as reconnaissance.

2) Politicians: We elect politicians (presidents, governors, legislators, etc) in part to be our moral leaders. They are supposed to take action based on what they believe is right or wrong and not based only on polls or their political party. Computers, and robots more generally, may certainly help our leaders. For example a robot may provide tele-presence for politicians so they may interact both with leaders at the capitol and community members at home. They may help calculate or simulate the possible effects of a given law or policy, and their exacting and tireless calculations are key in informing the politician.

Robots cannot, however, be moral leaders and therefore cannot take the place of a politician. A robot cannot write a law that dictates what is right or wrong in a certain jurisdiction, or what a fair penalty should be. A robot cannot choose to declare a state of emergency, inconveniencing some to potentially save others. A robot cannot decide whether to fund levees or other municipal projects that on one hand are expensive and use public resources, but could save lives or properties should a disaster strike. We ask our legislative and executive politicians to make these decisions, to weigh the good against the bad, and we place our trust in their moral leadership.

3) Judges and Juries: While robots can recount an entire code of laws in a fraction of a second, they cannot judge a case on its merits. One can conceive of an advanced computer program that when fed the transcript of a court case returns a ruling and sentence according to the laws in the current jurisdiction. However to test such a routine one would need to

compare its result to the consensus of several human judges, as any individual judge may deviate from case to case. Humans may not be perfect by this measure, however the notion of perfect is ill defined here. So long as they are honorably interpreting the laws and considering the merits of a case, we accept some variability in our judges. The role of the judiciary is to interpret the laws, not to compute them.

In many countries you have a right to be judged by a jury of your peers. The main reason is that if a jury is required to make a decision that can have such a great impact on someone's life, stripping them of their freedom and sending them to jail for example, it is important for the decision to be made by a person or persons who can empathize both with the accused and with the accuser. Different jurists will have different moral styles and therefore they may come up with different verdicts. It is not which particular verdict in a particular case that makes the jurist a moral actor, but the process of deciding right from wrong.

B. Uses of Autonomy

While the previous sections have shown that the use of robots for moral actions should be precluded, that does not mean that work on autonomous systems should be halted. While a robot and its algorithms cannot carry out a moral action on its own, it can certainly use its sensors, calculating abilities, and physical interactions to aid humans. Humans have always turned to technology to aid them in making decisions, and in caring out actions based on those decisions. With ever more accurate and capable sensors, and sophisticated algorithms to process that sensor information, a robot can give a soldier more situational awareness. It can tell a soldier that the enemy soldier went into a house, or that it saw something move that it thinks is a civilian. The moral hazard arises if a soldier begins to substitute the robot's judgment for his or her own on what to do in a certain situation.

In fact the inability for a robot to make moral decisions does not appear to restrict them from many roles in which humans are currently employed. Building and driving cars do not inherently require moral actions but rather consistent and precise execution of the robot's programming – indeed these are some places where robotics is making great strides. Machines are often used to clean our laundry and dishes, though we don't often use the term robot in this case. As technology progresses it is easy to imagine robots helping in construction, logistics, retail, and other professions where it is in fact preferable that they simply and reliably perform the same actions, and not change them based on moral hazards. There will certainly be other issues to consider, and likely some roles that humanity decides not to relinquish to a robot, but the inability to be a moral actor is only a hindrance where such morality is required.

IV. Conclusion

Robotics, like any emerging technology, raises new ethical questions that humanity must carefully examine. It is far better to consider these questions now, while the technology is in its infancy, and not wait until after its use in the world. Many of the scientists who worked on the atomic bomb later regretted their part in the creation of such a weapon [12], even though the fundamental math and science needed in its creation certainly had academic merit in their own right.

Autonomous robots with no human in the loop cannot be moral actors. They lack both the imagination to conceive of the effects should the principle of their actions be made universal, as well as the free will to make the choice to follow a moral style. There is no test of morality that a robot could pass as such as only the actions resulting from moral decisions are testable. They may appear to be acting morally, as they may take the same action we would expect a moral person to take, but that does not make them moral.

For these reasons they should not be employed in situations requiring moral action. They cannot be trusted to decide on killing humans, or on attacking buildings or vehicles, they should certainly have no autonomous lethal use. They are incapable of being moral leaders, and as such cannot replace humans in legislative, judicial, or executive governance. However that leaves ample territory where robots can help humans by doing what they are good at: exact computations, mechanical strength, and tireless focus.

References

[1] A. M. Johnson and S. Axinn, "The morality of autonomous robots," *Journal of Military Ethics*, vol. 12, no. 2, pp. 129–141, 2013.

[2] B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 177–190, 2003.

[3] R. Kirby, J. Forlizzi, and R. Simmons, "Affective social robots," *Robotics and Autonomous Systems*, vol. 58, p. 322, March 2010.

[4] A. F. Beavers, "Between angels and animals: The question of robot ethics, or is kantian moral agency desirable," in *Association for practical and professional ethics, eighteenth annual meeting, Cincinnati, Ohio, March*, 2009, pp. 5–8.

[5] V. Vinge, "The coming technological singularity," *Whole Earth Review*, vol. 81, pp. 88–95, 1993.

[6] I. Kant, *Opus Postumum (English Translation)*. Cambridge: Cambridge University Press, 1993, translated by Eckart Forster and Michael Rosen.

[7] S. Axinn, "Moral style," *The Journal of Value Inquiry*, vol. 24, no. 2, pp. 123–133, 1990.

[8] I. Kant, *Foundations of the Metaphysics of Morals*. New York: The Liberal Arts Press, 1959, translated by Lewis White Beck.

[9] M. L. Thompson, Ed., *Imagination in Kant's critical philosophy*. Berlin: De Gruyter, 2013.

[10] R. Arkin, *Governing lethal behavior in autonomous robots*. Boca Raton, FL: CRC Press, 2009.

[11] Department of the Army, "FM27-10. The law of land warfare," 1956, still authoritative. [Online]. Available: http://www.loc.gov/rr/frd/Military_Law/pdf/law_warfare-1956.pdf

[12] L. Szilard, "A petition to the president of the united states," July 1945. [Online]. Available: http://www.dannen.com/decision/45-07-17.html