Undergraduate Writing Colloquium Contest

Gabriel Darmon

Category: WRT (105E)

Creating Morality: The Ethical Implications of Intelligent Machines

The possibility of a machine that is as sophisticated and intelligent as a human being brings with it many complex ideas. While there are numerous aspects of an intelligent machine that could be considered, a rather intriguing area of artificial intelligence deals with morality. The technology available to humans is rapidly advancing and allows for machines to be increasingly autonomous. At a certain level of machine independence, one may wonder if artificial intelligence can be treated as a moral agent. If the point at which autonomous machines truly reflect human intelligence can be reached, should these machines have the potential to act morally? More significantly, would these moral machines deserve similar ethical treatment to that given to humans? Can a robot be held accountable for its actions? If an ethical science can be perfected, intelligent machines should be given a moral capacity similar to that of humans. Furthermore, humans should be able to hold robots morally responsible for their actions and in turn treat those robots as moral agents.

In order to understand the possibilities of moral machines, one must first analyze the reasons behind their necessity. It is no question that technology is a rapidly growing part of our lives. We are dependent on the benefits of machines daily. The basis of this dependence stems from the fact that technological growth is moving towards autonomy. The goals of technology revolve around the idea that humans should have to do as little as possible. These goals are realized by autonomous machines. The development of self-driving cars by Google represents the epiphany of this realization. In an article in *Financial Times*, the car's prototype is described as having the potential to "take over the world." Taking the extremely complex and dangerous task of driving and teaching a

machine to perform it without any human intervention demonstrates this idea of technological autonomy. However, this concept isn't restricted to driving.

The possibilities of machine independence are endless, and bring with them moral dilemmas to consider. A growing field in autonomous technologies deals with socially assistive robotics, or SAR (Feil-Seifer & Matarić, 2011). These machines, which primarily aid people with special needs, can impact the lives of children and the elderly significantly. Children with autism often benefit from these types of robotics through encouragement to "initiate and sustain social interaction" (Feil-Seifer & Matarić, 2011, p. 26). For instance, robots that can tell jokes to and play games with autistic children have the ability to ameliorate their social behaviors. However, the ethical concerns surrounding these machines are significant. A parent surely wouldn't leave their child in the care of a robot if it was unsure how it would behave morally. How can we ensure that a child is safe in the hands of a machine? In an article that analyzes this field of robotics, Feil-Seifer and Matarić (2011) define certain criteria for the autonomy, beneficence, and justice of these robotic machines. The fact that these areas need to be clearly explained shows that ethics must be considered whenever autonomy is a factor. Young children often develop strong connections with childcare figures. Taking advantage of this tendency by ensuring robots function morally provides a positive role model for growing children. Especially in childcare, machine independence and morality should not exist on their own.

The controversy of autonomous machines also extends to military applications. The popularity of using robots in war is growing. Currently, any armed, unmanned weapon has a "man in the loop," ensuring that any lethal action is approved by a human

(Altmann, Asaro, Sharkey & Sparrow; 2013). However, there is widespread concern about plans to remove this human intervention. It is certainly reasonable to worry when robots have lethal capabilities that are out of human control. Current drones aren't lacking in controversial performance. In early 2014, a US drone misidentified targets in Yemen and attacked a wedding procession, killing 12 civilians and wounding 15 others (*Turning a Wedding*, 2014). Obvious problems develop when it is realized that "no computational system can discriminate between combatants and innocents in a close-contact encounter" (Sharkey, 2008). The controversy over the possibility of these autonomous war machines is imminent. Controversy over whether or not war itself is ethical represents another debate. But how can we trust that a machine is performing at the ethical standard that we expect of it? This question can be answered if machines are programed to act morally.

This introduction of morality in the sphere of technology can be difficult to grasp. Humans have yet to map out the entirety of moral thinking and so integrating this into a rapidly growing scientific field seems out of place. After all, ethics can hardly be reduced to a science (Wallach, Allen & Smit; 2007). However, if scientists amount to the challenge of programming a human-like moral capacity into the functions of a machine, we can develop a significantly deeper understanding of human moral faculty. In order to implement this sort of complex behavior, ethical values must be broken down into their simplest forms. What is more significant is that the achievement of this feat allows machines that perform at a high degree of autonomy to also perform at a higher moral level. Essentially, autonomy and morality must be taken into account together. In the realm of machines, one should not exist without the other.

While it is clear that ethical functioning is a necessity for autonomous technology, the synthesis of the two may still be difficult to achieve. There are various speculations as to how this may be done. Colin Allen, Iva Smit, and Wendell Wallach analyze two different methods of creating artificial "sensitivity to the values, ethics, and legality of activities" (2006). They are referred to as "top-down" and "bottom-up" approaches. The top-down approach involves formulating a complex and extensive set of rules and algorithms based on known ethical and moral values. However, as Allen, Smit, and Wallach discuss, this approach brings with it many complications. While a strict set of rules or "commandments" may seem rather simple to execute, they appear to conflict with each other very easily. It isn't possible to accurately create rules for every possible scenario a robot may encounter. This poses a multitude of dilemmas for artificial intelligence attempting to act morally.

I feel it is necessary to avoid the first approach and consider the bottom-up solution in order for a machine to successfully evaluate an action morally. This approach resembles the education of a child. Children first learn moral values through social experiences. They can identify "appropriate and inappropriate behavior without necessarily providing an explicit theory of what counts as such" (Allen et al., 2006). We may not always be explicitly told what is right and wrong, but our moral values have developed over years of experience and are ingrained in our decision-making. If this method of learning and acquiring information is implemented in the engineering of autonomous machines, artificial morality can successfully be achieved. This is one of the most effective ways to allow humans to learn. When considering technology, machines must be allowed to "learn" moral concepts.

While the idea may seem abstract, the ability of machines to learn from their actions is certainly very possible. In a video on YouTube published by TED Talks, Raffaello D'Andrea demonstrates the athletic capabilities of quadcopters. An interesting part of the video shows D'Andrea tossing a ball to a quadcopter with a small racquet fastened to it, allowing it to strike the ball and send it back to the thrower. What is intriguing is that the quadcopter first misses the ball; but after a few throws, it is able to strike the ball back to D'Andrea with increasing accuracy. On the second throw, the quadcopter makes contact with the ball, but isn't able to strike it accurately, resulting in D'Andrea having difficulty catching it. However, by the third and fourth trials, the quadcopter is able to strike the ball with astounding accuracy. While this attests to the impressive athletic power of quadcopters specifically, it demonstrates something more meaningful: machines have the ability to learn. While the video by TED Talks restricts its demonstrations to physical actions, there is no reason why this ability cannot be extended to the sphere of morality and ethics. If a machine can learn from its experiences and independently acquire new information over time, the "bottom-up" approach to artificial morality can be considered. This learning is an essential approach to implementing an ethical capacity in technology.

The logical progression in consideration of moral machines is to study their interaction with humans. While the task of realizing artificial morality is difficult, yet another complex challenge arises when it is allowed to come into contact with actual human morality. How can the two coexist? How can a human justice system apply to non-human moral machines? A critical dilemma in this challenge is whether or not moral robots can be held accountable for their actions in the same way that humans are. I

believe the answer to this question is that we can treat moral machines as persons deserving moral treatment. However, there is some uncertainty in this conclusion.

This uncertainty arises when considering the point at which machines achieve this moral standing. How can we know that this level of artificial intelligence and ethical capability has been reached? Rob Sparrow describes a scenario known as the "Turing Triage Test" (2012, p. 301). The test is based on Allen Turing's famous "Turing Test," which he introduced in Computing Machinery and Intelligence to address the possibility of a machine convincingly imitating a human (Proudfoot, 2013). The new test is a scenario in which the subject must decide between sacrificing a human being and an intelligent machine. He believes that "machines will have achieved the moral status of persons" when this decision is truly a moral dilemma. If people have significant difficulty in making that decision, humanity will have crossed a threshold. Once this point is reached in the progression of artificial morality, further issues will arise regarding the treatment of moral machines. These issues involve punishment and emotion and their correlation to domesticated animals.

A common idea surrounding the treatment of these machines as moral agents is their comparison to pets and how they are punished. Punishment is an interesting aspect of ethics to consider when dealing with moral machines. Nancy G. Lin, Keith Abney, and George A. Bekey note a common counter argument to the belief that machines should be treated as moral persons. "My cat is not put on trial for arson when it knocks over a candle and burns down the house" (2012, p. 47). This is certainly true; it wouldn't make sense to treat a cat at the same moral level as a human. However, this is no reason that a machine with moral faculty shouldn't be treated at this level. The cognitive functions of a

cat are most likely very dissimilar to those of humans. While one may argue that machines don't "think" as humans do, the implementation of their ethical capacity would be performed and guided by humans. We have no control over the moral development of a cat, so it would be unreasonable to morally treat a cat as we treat humans. However, if we are studying and analyzing the possibility of realizing artificial morality in machines, this morality will surely resemble and reflect human values. For this reason, it would be logical to treat intelligent, moral robots as moral persons with similar status as humans. Artificial moral behavior would be held at a standard that humans defined.

This analogy to how animals are treated extends to emotion and its role in human-machine interaction. Piers Benn discusses in depth the reasons behind the feelings that we have towards people (1998). He explains in a chapter entitled, "Free will and the moral emotions," that we can only feel true moral anger "towards beings we take to be morally responsible for their behavior" (1998, p. 152). In terms of humanoid robots, we can only treat them morally when we can hold them morally responsible. This relates to human's treatment of animals. Dogs don't know better when they decide to pee on the floor of the living room. Dog owners can't truly be angry at their dogs for something like that because they know that dogs don't have the same moral faculties as humans. But a machine, with an implemented ethical code whose roots are based in the moral values that humans hold, can certainly be held responsible for its actions and if necessary, perhaps punished. While the idea of punishing and rewarding a machine seems abstract, it certainly doesn't signify that it isn't possible. Machines could be rewarded with richer data for behaving ethically (Allen et al., 2006). As fields and technologies advance, new terms and concepts develop alongside them.

It is certainly a new idea to consider, but the moral treatment of machines is imminent. Because the previously cited "bottom-up" approach is the most natural and effective way to implement artificial morality, we are relying on the capability of machines to learn. However, as Allen, Smit, and Wallach point out, "any system which has the ability to learn can also potentially undo any restraints built into the system" (2006, p. 153). This introduces a significant potential danger to humanity and only emphasizes the need to hold moral machines accountable for their actions.

John Basl best summarizes the dilemma of moral human-machine interaction: "being a moral patient is a function of the capacities an entity has, not the type of being that it is" (2012, p. 18). Does it matter that these agents are machines? We can ignore this fact assuming artificial intelligence has advanced sufficiently. Johnny Hartz Søraker likely agrees. Søraker believes that "a difference in treatment or value between two kinds of entities can only be justified on the basis of a relevant and significant difference between the two" (2012, p. 78). If machines reach the point of equal moral capacity to humans, than there is no significant difference between human ethics and artificial ethics. At that point, it would only be unjust to consider machines as morally lesser agents. At the moment, we have trouble believing that any current machine has genuine moral capabilities, let alone any true cognitive function. It is quite clear that technology and research have not yet advanced sufficiently. However, it is surely an attainable challenge in the future.

Artificial morality in machines likely has many possibilities that can't be imagined. Moral machines could prove to be useful in jobs and fields like law enforcement and childcare. However, these things can't be easily conceived until the

actual technology is perfected and realized. As long as technological autonomy advances, artificial morality must progress with it. Additionally, if this feat is accomplished, machines will deserve the same moral consideration that is given to humans.

References

Allen, C., Smit, I., Wallach, W. (2006). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7,* 149-155. doi:10.1007/s10676-006-0004-4

Altmann, J., Asaro, P., Sharkey, N., Sparrow., R. (2013). Armed military robots: editorial. *Ethics Inf Technol, 15,* 73-76. doi:10.1007/s10676-013-9218-1

Basl, J. (2012). Machines as Moral Patients We Shouldn't Care About (Yet): The Interests and Welfare of Current Machines. *The Machine Question: AI, Ethics and Moral Responsibility,* 17-24.

Benn, P. Ethics. London, GBR: Taylor and Francis, 1998. ProQuest ebrary. Web. 15 November 2014. Copyright © 1998. Taylor and Francis. All rights reserved.

D'Andrea, R. [TED]. (2013, June 11). *The astounding athletic power of quadcopters* [Video file]. Retrieved from https://www.youtube.com/watch?v=w2itwFJCgFQ

Feil-Seifer, D., Matarić, M. J. (2011). Socially Assistive Robotics: Ethical Issues Related to Technology. *IEEE Robotics & Automation Magazine*. doi:10.1109/MRA.2010.940150

Goodman, A., González, J. (2014). *Turning a Wedding Into a Funeral: U.S. Drone Strike in Yemen Killed as Many as 12 Civilians*. Retrieved from http://www.democracynow.org/2014/2/21/turning_a_wedding_into_a_funeral

Google car: Beep-beep. (2014). FT.Com, Retrieved from http://search.proquest.com/docview/1540875658?accountid=13567

Guarini, M. (2013). Introduction: Machine Ethics and the Ethics of Building Intelligent Machines. *Springer Science+Business Media Dordrecht, 32,* 213-215.doi:10.1007/s11245-013-9183-x

Lin, N. G., Abney, K., and Bekey, G. A., eds. Robot Ethics: The Ethical and Social Implications of Robotics. Cambridge, MA, USA: MIT Press, 2012. ProQuest ebrary. Web. 14 November 2014. Copyright © 2012. MIT Press. All rights reserved.

Miller, K., Wolf, M. J., Grodzinsky, F. (2014). Behind the mask: machine morality. *Journal of Experimental & Theoretical Artificial Intelligence,* 1-9. doi:10.1080/0952813X.2014.948315

Proudfoot, D. (2013). Rethinking Turing's Test. *The Journal of Philosophy, 110,* 391-411. doi:0022-362X/13/1007/391–411

Sharkey, N. (2008). The Ethical Frontiers of Robotics. *Science, 322,* 1800-1801. doi:10.1126/science.1164582

Søraker, J. H. (2012). Is there a continuity between man and machine? *The Machine Question: AI, Ethics and Moral Responsibility,* 78-82.

Wallach, W., Allen, C., Smit, I. (2007). Machine morality: bottom-up and top-down approaches for modeling human moral faculties. *AI & Soc, 22,* 565-582. doi:10.1007/s00146-007-0099-0