Leila Scola

Professor Vallor

Ethics in Technology

10 March 2018

Rational Artificial Intelligence

As modern technology continues to develop and improve, the human race is approaching the creation of beings that are able to mimic our cognitive abilities, raising several ethical questions about both the design and implementation of these mechanical beings. We are beginning to create machines that interact with us socially and perform tasks only capable of humans before. We must consider the ethical implications of advancing artificial intelligence (AI), and how we interact with them. Throughout the rest of the essay I will argue that cognitive and functional AI that lack emotions, and therefore desire, will most successfully ease human suffering, while these beings experience no emotion, therefore no suffering or desire. Using several different ethical theories, I will explain this design of AI should be the most ethical.

In order to fully appreciate the ethical magnitude of introducing emotions to robots, robots and artificial intelligence must be defined. The most simple definition of a robot is describing it as an embodied machine. AI on the other hand, is a type of programing that allows machines to perform tasks that formerly required intelligent humans. AI's currently possess calculative, predictive, pattern-recognition, and task-matching learning and intelligence. We currently have AI working in industry, health care, and the military, performing repetitive tasks that require little human interaction, and little skill in distinguishing objects. As technology advances, AI is beginning to gain these abilities, such as identifying facial expression and how to

respond. As robots reach these states, they begin to pervade society and human social institutions, posing questions of ethical significance. These questions include whether robots should be implemented with emotions as they approach the same cognitive and problem solving capabilities of humans, what rights robots should be given, and how we should treat robots. While they can interact socially with humans, they do not have true social intelligence. In order to be an effective AI, emotions are not required, just an ability to perform tasks that require intelligence well.

The current purpose of technology and robots is to ease human suffering and increase our productivity. We should not allow robots to have emotions because it would result in the creation of irrational, therefore unethical beings, that will compete with human desires and resources, while maintaining an intelligence ability and processing speed that could result in overpowering the human race if unchecked. As technology progresses, we must be aware that someday AI may have true social intelligence and the cognitive ability of humans, and we must consider the consequences of allowing AI to emote now. This description of AI, and those to come, are the machines I believe should not be given the ability to emote.

I believe that despite lacking emotions, rational AI will be able to make the same ethical decisions humans are capable of making. As expressed in Kantian ethics, an ethical being is one that does not act on desires, but rather out of duty and rationality (Kant, 4). A being that acts rationally is one that contains the highest morality, and therefore has dignity and worth. Currently, AI is able to rationally assess options for courses of action and choose ethically, without being swayed by emotions. A rational agent is one that "does the right thing … when an agent is plunked down in an environment, it generates a sequence of actions according to the

percepts it receives. This sequence of actions causes the environment to go through a sequence of states. If the sequence is desirable, then the agent has performed well"(Davis 2010, 36-37). Our goal is to make beings that are able to ease human suffering and aid us in our day to day tasks. If beings are rational they are able to do this to the best of their ability. They will not be swayed by emotions of anger, fear, desire, or bias. They are able to choose the right decision and understand a sequence of action. In addition, they will aid us in our tasks, easing our stress, because they possess no personal desires. If the AI was more emotional, it may act selfishly, or too hastily out of panic or fear of the consequences of it's inaction, rather than careful, rational assessment and a choice of the most rational, and ethical, response.

In addition, just because AI does not possess the ability to empathize, it can still rationally understand human suffering. Rational beings can understand the experience of others without having to feel it. As described by someone who cannot experience pain, yet understands the concept, " 'I have never doubted that pain is purely bio-physical in nature … Prior to my treatment pain was unknown to me in the sense that I had never undergone it myself, never actually felt it… Yet I insist that my factual knowledge of pain was nevertheless complete' " (Copeland 1993, 178). Rational AI would be able to understand the concept of pain and the importance to humans of preventing it, and would want to prevent it too. Because AI is not swayed by desires or emotions of its own, it's only goal is to minimize human suffering through our commands. According to utilitarian framework of ethics, any action performed must maximize total pleasure rather than total suffering (Mill, 1). Under these principles, AI satisfies utilitarian ethics framework, as they minimize human suffering, and cannot experience suffering of their own. I therefore believe, rational AI, lacking the ability to emote, is more than capable of

making extremely ethical decisions that both follow the framework of Kantian and Utilitarian Ethics (Kant, 3; Mill, 1).

As mentioned previously, the goal of AI is to create beings that ease human suffering, and to create autonomous, emotional beings will only increase human stress, as well as AI stress, which is unethical for both the suffering humans and newly created suffering robots according to utilitarian points of view. AI have been created for the past couple of decades to work in industry to speed up manufacturing production and lower costs. AI is also being introduced into healthcare practices to keep the elderly or disabled on schedules and aid in their day to day tasks. According to Arnold, "The gradual development of complex machines and the gradual mechanization of work helped people provide necessities; these changes gave people more time to improve society" (Arnold 1986, 158). The improvement of society has been extremely ethical and useful. It has allowed humans to leave more time for higher learning and pursuing interests, both decreasing stress and suffering, while increasing pleasure. The point of AI is "to bring speed and reliability" which could not happen if the AI was overwhelmed with the same desires, interests, and emotions that humans possess (Kraus 2018, 1). The point of AI is to ease human suffering and perform the tasks we do not wish to, therefore it makes no sense to design intelligent beings with the same cognitive abilities as humans and the same ability to suffer. In essence, we would simply be recreating human beings. Humans are already "'predictably irrational'" and prone to mistakes due to their emotions(Davis 2010, 619). Humans "sometimes used their intelligence in aggressive ways because humans have some innately aggressive tendencies, due to natural selection. The machines we build need not be innately aggressive"(Davis 2010, 1037). AI would never make aggressive or unethical decisions due to an

outpour of emotions. We want to create rational beings that minimize irrational behavior, as they are are not prone to outbursts that often instigates mistakes and unethical behavior. It is imperative that if we are to create ethical beings that "AI system must function with far greater reliability than a set of experts" as "humans know that humans make mistakes" (Davis 2010, 166). Creating AI without emotions will decrease human stress and suffering, while avoiding the suffering of AI that have the potential to act as unethically and suffer as much as human beings.

According to Buddhist and Utilitarian framework of ethics, it is immoral to create emotional AI due to the suffering they will inevitably face. Buddhist believe that to create any suffering is immoral and therefore unethical (Lin 2014, 70). Utilitarians believe that in order to act ethically, one must always maximize the amount of happiness and decrease the amount of suffering with their actions in order to act ethically (Wall, 31). Introducing AI that have the ability to suffer is innately unethical. In addition, AI will then have the ability to inflict suffering onto human beings. If AI were rational, their only utility would be to ease human suffering, acting as tools and easing mechanical workloads. Despite emotional AI's ability to intermittently feel happiness, the amount of total suffering would outweigh total happiness, as both AI and humans would suffer. If the AI was simply rational, it would be able to ease human workload, and therefore suffering, while providing humans with time for cultivation of virtues and other pleasures. This means that rational AI only possess the ability to cultivate happiness, and will never suffer, or intentionally inflict suffering. Rational AI can only increase total pleasure. In contrast, emotional AI can temporarily increase pleasure, but will ultimately introduce a large amount of total suffering. Buddhist and Utilitarian framework of ethics therefore is in support of the creation of rational AI versus emotional AI.

The creation of emotional AI would lead to another race of beings that compete for human resources, compounding suffering on both races' accounts. Joanna Bryson argues that it would be unethical to to create AI that compete with us, "to make them suffer, or to make them unnecessarily mortal" (Bryson 2018, 22). She argues that agents with desires and intelligence would soon turn to actions that would satisfy their own personal desires, not humans (Bryson 2018, 26). She questions, "Would an initially-human-like capacity for computation be worth sacrificing human potential for in order to create something eventually as similar to us as crabgrass (Moore 1947)?" (Bryson 2018, 26). We would sacrifice our own intelligence and resources to create AI. She states that it would be wise to think of AI as an extension of our "own motivational systems" so that they can maintain our situation (Bryson 2018, 29). This is the most ethical route, as it both increases our happiness and keeps these agents from suffering, or turning to their own desires and inflicting suffering on the human race as we compete for our desires and resources.

There is a certain hubris in believing we can not only create AI with emotions, but we can create emotions that are perfectly well designed. We know human emotions are unpredictable and complicated, and robot emotions would be too. Human emotions are unpredictable, irrational, and often lead to our hamartias. The purpose of the development of AI was to reduce human error and unpredictability. We have purposely created rational beings in order to reduce the mistakes created by human outbursts and irrationality. The introduction of emotional AI would not only be disastrous, but unethical. Kantians believe the most moral decisions are those made in rationality, foregoing desires and emotional wants (Kant, 3). Utilitarians believe ethical decisions minimize suffering (Mill, 1). Rational AI both forego desires and minimize suffering,

while reducing human irrationality and suffering. Not only is it the most ethical decision to create rational AI that may have unpredictable emotions, but the creation of AI with unpredictable emotions would lead to selfish AI as mentioned previously. It is naive to believe emotional AI will continue to fulfill human desires when they have their own. Not only will AI act on their own desires, which has been proven to be unethical, but they have the potential to manipulate human desires. With the creation of beings theoretically more intelligent than humans with faster processing speeds, and emotions, their ability to act on their own desires is magnified. They possess the ability to monitoring and manipulation of users' emotions by artificial systems" (Lin 2014, 236). The creation of emotional AI would reduce rationality further and increase suffering, which is unethical and undesirable.

A defender of AI emotion may argue that without implementing emotions, AI will never be able to empathize with humans or our common goal of a beneficial society, and therefore will never be able to act morally (Thagard 2017). Without being able to understand human suffering, AI will not be able to understand the importance of preventing it. If AI do not understand suffering, they may never be able to fully realize when they are creating suffering, and therefore, will often act unethically according to a utilitarian framework of ethics (Mill, 1). Paul Thagard philosophizes,

> "according to obsolete ideas, rationality and emotion are fundamentally opposed… But there is abundant evidence … that cognition and emotion are intertwined in the human mind and brain. Although there are cases where emotions make people irrational… Emotions help people to decide what is important and to integrate complex information

into crucial decisions….if robots are going to be ethical in the way that people are, they need emotions" (Thagard 2017).

The inconsistencies of life make it so not every situation can be predicted and dealt with purely rational means. AI must be able to empathize with human situations so they may understand the value of the suffering a human may be facing if certain actions, while more rational and effective, are ultimately more hurtful than taking a less effective direction to solve a problem. For example, if a human has their leg pinned under a car, the rational AI may believe that it is easier to cut off a human's leg, since it could not be repaired even if removed. This uses rational logic and is a quick solution to the problem. However, another human being with empathy for the trapped human would recognize the fear and pain involved in having a leg amputated, and would instead wait to have the car lifted in case there is a chance of repairing the leg, despite this taking more time. Without emotions, robots cannot understand the most helpful route for humans to truly minimize suffering, and therefore will often act unethically.

I disagree with the logic of those that believe emotions should be implemented in AI. If an AI was placed in the aforementioned scenario, rational AI is intelligent enough to understand the human's leg has a chance of being saved. It would not react emotionally and attempt to save the human without thinking through multiple routes of action. The AI would rather wait for more assistance and be able to properly assess the situation and help the human to the best of its ability. The rational AI can calmly understand the situation, process multiple routes of action, decide on the best one, and enact without irrational behaviors, such as panic, in turn making hasty or clumsy mistakes. The rational AI will be able to truly minimize human suffering, continuously acting ethically.

Advocates of implementing AI emotions may claim that in addition to being unable to comprehend human suffering, rational AI would not be able to understand human affection and will negatively impact human relationships. Humans have a tendency to anthropomorphize many objects, including machines. This leads to the belief of a real connection to devices that cannot ever empathize or comprehend the qualities being attributed to it. The ease of interacting with helpful AI allows humans to become attached to these devices and in some cases even prefer to interact with AI rather than humans. The AI is helpful, always available, and does the exact bidding of the user. It can be argued that while this is helpful, it is not ethical. Positive virtues, such as courage, empathy, and wisdom, can only developed through empathetic and charged interactions. As explained by Shannon Vallor human beings, "unlike machines, live as members of a complex, diverse and overlapping social collectives that require them to develop and constantly negotiate among various holistic conceptions of value" (Vallor 2017, 173). Humans constantly need to consciously put effort into practicing these techniques in order to develop them and become a well rounded moral agent. On the other hand, "intelligent machines… lack the distinctive social and psychological conditions that make wisdom possible" (Vallor 2017, 174). Machines do not possess the ability to cultivate the virtues that build the backbone of our holistic society. In order to maintain moral agency, we need machines that can interact with us in the same cognitive and emotional ways we interact with each other. From the standpoint of a virtue ethicist, the introduction of rational AI into is extremely unethical (Aristotle, 3). In order to continue cultivating these positive virtues, emotional AI that possess moral agency must be created. In order to create ethical AI, emotions must be implemented to develop the same virtues in AI and retain them in humans, creating more dignified and happy beings.

Despite this objection to my argument, I believe the introduction of rational AI into society will only be unethical if we allow ourselves to become so reliant on AI that we no longer have positive human interactions. The issue with rational AI is not that it inherently blocks the cultivation of virtues, but that we become so reliant we forego the negative human interactions that cultivate wisdom and other virtues in favor of the convenient relationships AI provides. Virtues require many opportunities to practice them in order to be developed and retained. While emotional and empathetic AI would be able to aid in our cultivation of virtues, I believe they will also have the ability to inflict suffering, rather than solely positive and helpful interactions, which is therefore unethical in a utilitarian framework of ethics (Mill, 1). As long as humans remember rational AI should be used as a tool to aid in our daily lives in order to create more time to practice virtues, we may retain our virtues, decrease total suffering, and forego creating more irrational beings motivated by desire. We avoid the overwhelming unethical affects, and still boost moral agency of humans and their ethical decisions. Rational AI will ultimately be the most ethical choice for humans as long as we can continue to use them wisely.

Artificial Intelligence was created in order to reduce human stress and suffering, and this status quo can only be maintained if AI continues to lack emotion and remains rational. Implementing emotion decreases the rationality of AI and cause it to act unethically when making emotional decisions. Rational AI hold the ability to analyze and respond appropriately to situations without being swayed by emotions, which could cause AI to act hastily or in biased ways. In addition, rational AI has been created with the purpose of easing human suffering and workloads but decreasing human error and stress. Introducing emotional AI would only compound the same suffering and workloads that emotional humans create for themselves. Not

only would it compound it, but this would lead to a competition for resources as emotional AI pursue their own desires. This could be disastrous as these intelligent beings could learn to manipulate us to fulfill their own desires. Overall, introducing emotional AI would increase total suffering, create more ebings driven by desire, and decrease the cultivation of virtues, unethical in the framework of utilitarian, kantian, and virtue ethics, respectively(Mill, 1; Kant, 3; Aristotle, 5). Rational AI the only way to create ethical beings that retain the ethical framework of our society.

Reference List

Aristotle. *Nicomachean Ethics.* Excerpts in class handout.

Arnold, William R., and John S. Bowie. 1986. *Artificial Intelligence: A Person, Commonsense Journey.* Englewood Cliffs: Prentice-Hall Inc.

Bryson, Joanna J. 2018. "Patiency is not a Virtue: the Design of Intelligent Systems and Systems of Ethics." *Ethics and Information Technology* 20, no. 1 (Winter): 15-26. https://link.springer.com/article/10.1007/s10676-018-9448-6

Copeland, Jack. 1993. *Artificial Intelligence: A Philosophical Introduction.* Oxford: Blackwell Publishers.

Davis, Ernest. 2010. "Intelligent Agents." *Artificial Intelligence: A Modern Approach,* edited by Stuart Russell and Peter Norvig, 34-59. Upper Saddle River: Pearson Education, Inc.

Kant, Immanuel. *Groundwork for the Metaphysics of Moral.* Excerpts in class handout.

Kraus, Jean-Louis. 2018. "Is Artificial Intelligence Associated with Chemist's Creativity Represents a Threat to Humanity?" *AI & Society* 146, no. 1 (Winter): 1-3.

Lin, Patrick, ed. 2014. *Robot Ethics: The Ethical and Social Implications of Robotics.* Cambridge: The MIT Press.

Mill, John Stuart. *Utilitarianism.* Excerpts in class handout.

Thagard, Paul. 2017. "Will Robots Ever Have Emotions?" *Psychology Today,* December 14, 2017. https://www.psychologytoday.com/blog/hot-thought/201712/will-robots-ever-have-emotions

Vallor, Shannon. 2017. "AI And The Automation Of Wisdom." In *Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics*, edited by Powers, Thomas, 161-178. New York City: Springer International Publishing.

Wall, Thomas F. 2003. *Thinking Critically About Moral Problems.* Belmont: Wadsworth Group.