Moral Machines: Mindless Morality and its Legal Implications

Andrew Schmelzer

We live in an era of increasing automation. As we employ more robots, they will cause more morally harmful outcomes (Danaher, 2016). Artificial moral agency will relieve some of the moral, legal, and social strains the increasing use of robots places on society. Federal and international legal preparation and intervention can further eliminate the issues robotization causes.

What is morality?

We need a working definition of morality to consider programming morality into machines. DeBaets (2014) defines a choice or action as moral if it is intentionally oriented toward the good. An agent is moral if its actions are good. Despite differing stances, our current ethical systems "generally understand morality fundamentally as a necessary restraint on one's desire with the effect of, though not always for the sake of, promoting liberty and the public good" (Beavers, 2011). Morality does not apply to entities that cannot choose: we can be moral only because we can make decisions (Allen, 2011). By increasing the decision-autonomy of machines, we necessitate their morality.

We must implement morality in machines to prevent and reduce morally wrong outcomes caused by machines. Allen, Varner, and Zinser assert machines with autonomous capacity to do good also have capacity to do harm (2000). Our science fiction depicts futures where machines unbounded by morality cause immeasurable harm to humanity. While these images are entertaining and appalling, I focus more on the near and practical future. Allen proposes that the current reason for developing moral machines is to "forestall inflexible, ethically-blind technologies from propagating," and to make "machines whose controls involve increasing degrees of sensitivity to things that matter ethically" (Allen, 2011). "Ethically-blind" describes a system that does not determine its actions based on the moral and ethical circumstances it operates in. Allen (2011) gives the example of a car that requires breathalyzer input to start: it has rigid morality by design, but it cannot tell that an injured child requiring medical attention sits in the back seat. I assert that artificial moral agency serves to reduce the number of morally callous, wrong, and harmful outcomes that robotization brings.

Not all of our devices need moral agency. As autonomy increases, morality becomes more necessary in robots, but the reverse also holds. Machines with little autonomy need less ethical sensitivity. A refrigerator need not decide if the amount someone eats is healthy, and limit access accordingly. In fact, that fridge would infringe on human autonomy. Hellstrom (2013) discusses how autonomy comes in a gradient, from devices with little to no choices (such as this refrigerator), to autonomous units — such as self-driving cars with control over their direction, speed, and route, — to highly autonomous machines like the robotic companions we see in science fiction. A teleoperated machine has little to no autonomy because it has no decision making power; a human controls its actions (Hellstrom, 2013). The greater potential for good that a machine wields, the greater potential for harm it carries, and thus the greater need for ethical sensitivity.

Ethical sensitivity does not require moral perfection. I do not expect morally perfect decisions from machines. In fact, because humans are morally imperfect, we cannot base moral perfection off of humanity by holding machines to human ideals. Our moral development continues today, and I believe may never finish. Designing an artificial moral agent bound by the morality of today dooms it to obsolescence: ethical decisions from a hundred years ago look much more racist, sexist, etc., and less 'good' from today's perspective; today's ethics might have the same bias when viewed from the future (Creighton, 2016). Because the nature of our ethics changes, an agent will stumble eventually. Instead, we strive for morally human (or even better than human) decisions from machines. When a machine's actions reflect those of a human, we will have met the standards for artificial moral agency.

We can test for artificial moral agency with the Moral Turing Test (Allen, Varner, & Zinser, 2000). In the MTT, where a judge tries to differentiate between a machine and a person by their moral actions. An agent passes the test when the judge cannot correctly identify the machine more often than chance. Then, a machine qualifies as a moral agent. In the comparative Moral Turing Test (cMTT), the judge compares the behaviors of the two subjects, and determines which action is morally better than the other (Allen, Varner, & Zinser, 2000). When a machine's behavior consistently scores morally preferable to a human's behavior, then either the agent will have surpassed human standards, or the human's behavior markedly strays from those standards.

Thus far, I have defined a general form of morality for machine behavior: promoting liberty and the public good. I then covered why we require moral sensitivity and postulated which machines need ethical capacities. I stipulated ethical standards of an artificial moral agent and proposed a testing method of morality. Knowing what we need, I move on to how to design an artificial moral agent.

Developing Artificial Moral Agency

Top-Down Approaches

The first approach to developing artificial moral agency come from Isaac Asimov and his Three Laws of Robotics (1950):

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The ensuing fiction Asimov wrote based on these laws explores how they break down, failing their original purpose of maintaining order. The laws form the first attempt of a top-down approach to machine morality. The top-down approach creates a moral system ready for use and implementation with all features and functions included. Some of the methods classified as top-down include rule-based, virtue-based, and consequence-based systems (Allen, Varner, & Zinser, 2000).

Rule-based systems use lists of rules to follow. Some rules hold priority over other rules, and some situations invalidate or circumvent other rules when specific conditions occur. Theoretically, with enough rules and caveats, a machine can behave according to our moral standards, or even the standards we aspire to meet. However, rule-based systems break down when rules change, or when a rule contradicts itself. Furthermore, ethicists still deeply disagree about which ethical standards to base our machines on (Allen, Varner, & Zinser, 2000).

Virtue-based systems work by matching behavior to a list of virtues a machine ought to have. Frankena (1973) provides a list of terminal values — virtues that are valued for themselves, rather than their consequences (Yudkowsky, 2011):

> Life, consciousness, and activity; health and strength; pleasures and satisfactions of all or certain kinds; happiness, beatitude, contentment, etc.; truth; knowledge and true opinions of various kinds, understanding, wisdom; beauty, harmony, proportion in objects contemplated; aesthetic experience; morally good dispositions or virtues; mutual affection, love, friendship, cooperation; just distribution of goods and evils; harmony and proportion in one's own life; power and experiences of achievement; self-expression; freedom; peace, security; adventure and novelty; and good reputation, honor, esteem, etc.

Programming all of those values directly into a single utility function (the method of determining positive or negative results) is ridiculous. Can engineers or ethicists quantify each value and agree on a prioritization for each? Yudkowsky (2011) proposes a 'one-wrong-number' problem: a phone number has 10 digits, but dialing one wrong number does not mean you will connect with someone 90% like the person intended. The same may apply to virtue-based machines. Furthermore, some values we deem worthy of implementation in our machines may contradict each other, such as compassion and honesty (e.g. a child's professional baseball potential). In this way virtue-based systems still require the caveats of a rule-based system (Allen, Varner, & Zinser, 2000). But what about non-terminal virtues, that is, virtues we value for their repercussions?

Consequence-based systems work by evaluating the potential consequences of their behaviors. The machine determines the best method of action by determining how the action will affect the others in the situation. However, calculating every effect on every person in the scope of an operation can "be a computational black hole" (Allen, Varner, & Zinser, 2000). Calculating every action's effects on every individual involved for as long as those effects persist, along with the comprehensive scientific understanding required for such an evaluation could lead to computing indefinitely for a basic decision (Allen, Varner, & Zinser, 2000). However, we do not hold humans to that kind of standard of calculation, and methods to abbreviate such a calculation exist, such as intermediate goals and working backwards from a desired outcome (Allen, Varner, & Zinser, 2000).

That concludes the discussion of top-down approaches to machine morality. Rule-based systems require incredible complexity due to the flexibility of our practiced rules, but show great potential. Virtue-based systems may contradict themselves with the virtues we program in. Machines that calculate consequences of their actions may drive themselves computationally crazy, but we do have methods to quicken calculations. Next we explore the other approaches of design, bottom-up methods.

Bottom-Up Approaches

Bottom-up systems develop through experience. Instead of programming the behavioral guidelines directly, bottom-up systems mimic patterns of behavior deemed 'good' and cease behaviors deemed 'bad.' The three methods of bottom-up development I will discuss here are neural network learning, genetic algorithms, and scenario analysis systems.

Neural networks function similarly to neurons: connections between inputs and outputs make up a system that can learn to do various things, from playing computer games to running bipedally in a simulation. By using that learning capability on ethical endeavors, a moral machine begins to develop. From reinforcement of positive behaviors and penalty of negative ones, the algorithm learns the pattern of our moral systems. Eventually, engineers place the algorithm in charge of a physical machine, and away it goes. One downside to this is the uncertainty regarding what the algorithm learned. When the army tried to get a neural net to recognize tanks hidden in trees, what looked like a distinction between trees, tanks, and partly concealed tanks turned out to be a distinction between a sunny and cloudy day (Dreyfus & Dreyfus, 1992). Kuang (2017) writes about Darrel's potential solution: having two neural networks working side by side. The first learns the correlation between input and output, challenging situation and ethically right decision, respectively. The second algorithm focuses on learning language and connects tags or captions from an input and explains what cues and ideas the second algorithm used to come up with a course of action. The second weak point stems from allowing mistakes: no amount of learning can verify that the machine will act morally in all situations in the real world, including those not tested and learned from.

Genetic algorithms operate on a somewhat similar principle. Large numbers of simple digital agents run through ethically challenging simulations. The ones that return the best scores get "mated" with each other, blending code with a few randomizations, and then the test runs again (Fox, 2009). After the best (or acceptably best) scores based on desired outcomes are achieved, a new situation is added to the repertoire that each program must surpass. In this way, machines can learn our moral patterns. Once thoroughly evolved, we implement the program, and the machine operates independently in the real world. Alternatively to direct implementation, we could evolve the program to learn patterns quickly and efficiently, and then run it through the neural network training. This method suffers the same downsides as neural networking: we cannot tell what it learned or whether it will make mistakes in the future.

The final approach involves scenario analysis. Parkin (2017) describes a method of teaching AI by having it read books and stories and learn the literature's ideas and social norms. While this may not apply to machines we do not intend to behave as humans, the idea still applies to niche or domain-specific machines. Instead of using literature as learning input, we provide a learning program with records of past wrongdoings and successful outcomes of ethically-blind machines in its niche. Then the program could infer the proper behaviors for real world events it may encounter in the future. After analyzing the settings and events of each scenario, the program would save the connections it made for later human inspection. If the program's connections proved 'good,' it would then receive a new batch of scenarios to test through, and repeat the cycle. One downside to this approach involves painstaking human analysis. A new program would have to go through this cycle for every machine niche that requires a moral agent, and a human evaluator would have to carefully examine every connection and correlation the program develops. Darrel's (2017) explaining neural net could work in tandem with a scenario analysis system to alleviate the human requirement for analysis. This approach does get closer to solving the issue of working in new environments than the previous two approaches, but may nonetheless stumble once implemented in reality.

Bottom-up approaches utilize continued development to reach an approximation of moral standards. Neural networks develop connections and correlations to create a new output, but we struggle to know why the system comes to a decision. Genetic algorithms refine themselves by duplicating the most successful code into the next generation of programs, with a little mutation for adaptation. A genetic algorithm's learning process also remains obscured without careful record of iterations, which may be beyond human comprehension. Scenario analysis systems can learn the best conduct historically shown as ethically right, but still retains potential for error. As of yet, we do not have a reliable method to develop an artificial moral agent.

Allen, Varner, and Zinser (2000) propose a hybrid approach, meshing a top-down system with a bottom-up system to get the best results. I propose using a 3-in-1 machine hybrid: a rule-based system, a neural network, and Darrel's (2017) explaining algorithm. The rule-based system would handle much of the ethical work a machine encounters. The neural network supports the rule-based system by taking over when the rule system self-contradicts or runs into a loop error. Then an explaining algorithm can consider the situations and record the neural net's decision making process into storage. When such a machine errs, humans can open up the storage and figure out how the machine decided to act wrongly. Finally, humans can modify the ruleset given to the first system to account for the situations outside the ruleset's domain. My proposed system does, however, rely on a well devised ethical ruleset to continue functioning, and frequent updates to systems in operation. One could argue it combines the disadvantages of both systems, but I assert that a well-designed system-trio would prevent rule-based system breakdown, protect a wide domain of situations from neural network error, and be transparent in its failures, minimizing those downsides.

We have many approaches to develop machine morality. Each can fail in its own way, but through further research and combination of the methods discussed above, we may come to an implementable solution in the near future. Next, let us consider what delineates an artificial moral agent.

Requirements for Artificial Moral Agency

We now have a picture of how to design a robot that can follow the rules we impose on it, but we need to figure out what we need from a moral machine. To build an artificial moral agent, DeBaets (2014) argues that a machine must have embodiment, learning, teleology toward the good, and empathy.

DeBaets (2014) claims that moral functioning requires embodiment because if a machine acts in and influences the physical world, it must have a physical manifestation. "Embodiment [requires] that a particular decision-making entity be intricately linked to a particular concrete action; morality cannot solely be virtual if it is to be real" (DeBaets, 2014). They can work from a distance, have multiple centers of action, and have distributed decision-making centers, but each requires physical form. Embodiment constrains machines to a physical form, so this definition of moral agency excludes algorithms and programs that do not interact with the physical world.

Ethical machines need learning capacity so they can perform as taught by ethical and moral rules and extrapolate what they have learned into new situations. This requirement excludes any top-down approach that does not involve frequent patches and updates. Hybrid systems combine rule sets and learning capacities, and so fulfill this requirement since they can adjust to new inputs and refine their moral behavior. A stronger hybrid could also refine the base rulesets it has, but that may return to the uncertainty of a neural network's learned material discussed earlier. This leads us into the last two requirements.

Teleology toward the good and empathy both face a sizable complication: they both require some form of consciousness. For a machine to empathize with and understand emotions of others, it must have emotion itself. Coeckelbergh (2010) claims that true emotion requires consciousness and mental states in both cognitivist theory and feeling theory. Thus, if robots do not have consciousness or mental states, they cannot have emotions and therefore cannot have moral agency. Additionally, if a machine innately desires to do good, it must have some form of inner thoughts or feeling that it is indeed doing good, so teleology also requires consciousness or mental states. Much of human responsibility and moral agency relies on this theory of mind. In court, the insanity or state of mind defense can counter criminal charges. However, no empirical way to test for state of mind or consciousness in people exists today. Why require those immeasurable characteristics in our robots?

Emotionless Machinery

We interpret other humans' behaviors as coming from or influenced by emotion, but we have no way to truly determine emotional state. Verbal and nonverbal cues give us insights to emotions others feel. They may imitate or fake those cues, but we interact with them just the same as long as they maintain their deception (Goffman, 1956). We measure other people by their display or performance of emotion.

Since the appearance of emotion in people regulates social interaction and human morality, we must judge robots by that same appearance. Even today, machines can read breathing and heart rate (Gent, 2016), and computers do not need to see an entire face to determine emotion displayed (Wegrzyn, Vogt, Kireclioglu, Schneider, & Kissler, 2017). Soon enough, a machine could learn to display human emotion by imitating the cues they're designed to measure. In theory, a robot could imitate or fake emotional cues as well as humans display them naturally. People already tend to anthropomorphize robots, empathize with them, and interpret their behavior as emotional (Turkle, 2011). For consistency in the way we treat human display of emotion and interpret it as real, we must also treat robotic display of emotion as real. If the requirement for empathy changes from true emotion to functional emotion — as is consistent with how we treat people — then an imitating robot fulfills all the requirements for empathy, effectively avoiding the issue regarding consciousness and mental state. Compassion could be the reason an autonomous car veers into a tree rather than a line of children, but the appearance of compassion could also serve the same effect.

Additionally, a robot can have an artificial teleology towards good, granted that all of the taught responses programmed into the machine are 'good.' Beavers' (2011) discussion of classical utilitarianism, referencing Mill (1979), claims that acting good is the same as being good. The same applies to humans, as far as we can tell from the outside. Wallach and Allen (2009) note that "values that emerge through the bottom-up development of a system reflect the specific causal determinates of a system's behavior". In other words, a 'good' and 'moral' robot is one that takes moral and good actions. Thus, while we may not get true teleology, functional teleology can suffice.

We can fulfill all of DeBaets' requirements for an artificial agent. We have embodied machines. Our machines can adapt to their environments and learn relative to their situations. By considering our machines in the same way we consider humans, functional empathy and simulated teleology can fill true empathy and teleology. I agree with DeBaets that "[nothing] said here requires qualitative leaps in technological development over what already exists (DeBaets, 2014)." While all of these traits form a basis for artificial moral agents, even moral robotization will still cause problems in current social systems. Next, we discuss those problems.

Legal Issues of Robotization

Responsibility Gaps

Because of the autonomy gradient, responsibility gaps will arise. In today's world, when a robot harms someone, we look to the manufacturer or programmer for the three types of responsibility: causal responsibility, moral/legal responsibility, and liability responsibility (Danaher, 2016). In most cases, responsibility falls to the manufacturer or programmer for all three. However, as robotization increases, potential gaps in these responsibilities arise.

Danaher (2016) claims that as robotization increases, the amount of harm machines cause will also increase. Causal responsibility will always remain with the robot: if its actions harmed someone, it will have causal responsibility. Outside factors could force an ethically unprepared robot into a situation it cannot handle, making the encounter into an accident. Even in today's legal system, accidents cause hiccups and difficulties. Causal scapegoating through apparent "accidents" could occur with a moral machine as well.

If humans do not properly equip robots with moral capabilities, can blame fall on it for a moral error? We could fix this moral incompetence by giving all of our robots some form of moral software. Seeing as morality adapts over time, a robot will blunder eventually; no machine can have moral perfection because human morals change. When a robot harms someone, does moral responsibility fall on the engineers that designed the adaptive morality drive within the machine? If they could not predict the robot's behavior, they cannot shoulder moral blame

(Matthias, 2004). We must address this gap before a morally-independent machine harms someone and no one can take moral responsibility.

Liability responsibility may also drift off into a grey area. When a robot autonomous enough to develop and adapt to its situation harms someone, who does legal responsibility fall on? That is, who atones for the damage done by the robot? If the engineers had no way to predict the robot’s behavior due to the adaptive, learning design, they cannot hold liability responsibility. The robot itself may not understand what it has done or have the capacity to compensate the victim. Additionally, how do we make sure that no one can exploit this sort of grey area for their own machinations? We must build thorough legal and social systems to avoid liability evasion or scapegoating and insure against lack of liability.

The final gap I will discuss is the retribution gap. Humans innately seek retributive punishment (Danaher, 2016). An agent must have culpability for a victim to seek retribution towards the agent. Culpability, or capacity to receive blame, comes from both causal responsibility and moral responsibility (Danaher, 2016). One facet of this issue pertains to intention: can a robot have a desired outcome of its actions? Relying on the lack of consciousness idea above, it cannot. We treat accidental damage different from intentional harm in cases of humans, but a robot cannot intentionally harm without having intention. Another facet comes from the social aspect of retribution. Where do people go when they cannot fully take vengeance or receive proper compensation for harm? Simply having autonomous robots in the world puts everyone at risk for accidental damages.

Multiple responsibility gaps emerge with the furthered use of robotic technologies. Artificial moral agency alone will not fix these issues, but can reduce the rate at which they emerge. Still more issues stem from the extended use of robots in today’s world.

Other Robotization Issues

Computers are hackable. Despite all efforts to secure systems and data, hackers can and will exploit bugs and holes to serve their own purposes. A hacker can easily turn an AI weapon into an indiscriminate slaughtering machine (Musgrave & Roberts, 2015). The more robots we employ in daily life, the easier someone can abduct and modify one. Wilson illustrates the incredible danger of reprogrammed autonomous machines in *Robopocalypse* (2011). As we continue to develop better security measures, we may reduce the risk of illegal takeover, but proper precautions must be taken.

While automation and mass production can and has greatly reduced production costs in manufacturing (Roy, 2017), machines taking the role of humans in the production cycle causes economic hardship on the people displaced. Structural unemployment will only grow with further robotization (Brynjolfsson & McAfee, 2011). Increasingly advanced autonomous machines will make human participation in many of today's jobs obsolete and inefficient. Papps and Winkelmann (1999) show that increased unemployment also increases crime rates. Therefore, increased automation will increase crime rates unless we can employ displaced workers.

These issues can only grow larger the longer we procrastinate confronting them. Robots will continue to cause harm. Blame for damages will continue to fall on those not responsible for those damages. Those damaged will not have compensation from appropriate parties. People may not find proper targets for their retributive blame. Criminals could employ autonomous units for devastating nefarious schemes. People continue to lose their jobs as we discover new automation techniques and refine existing ones. We need legislation to assist with these issues that artificial moral agency cannot fix alone.

Legislation Complications

Developing good legislation will be crucial to guiding and controlling the development of robotics technologies. Leenes, Palmerini, Koops, Bertolini, Salvini, and Lucivero (2017) discuss several issues in creating and maintaining accurate, effective rulings. Good legislation faces three major pitfalls: technology-neutrality versus legal certainty, preemptive versus overdue legislation, and adjusting rulings for new technologies.

Technology-neutrality refers to the wide spread of the law. A ruling is technology-neutral if it applies to multiple technologies or forms an umbrella over a realm of technologies. Leenes, Palmerini, Koops, Bertolini, Salvini, and Lucivero (2017) use the example of ‘secrecy of communication’ versus ‘secrecy of the post’ to clarify the difference in specificity. Technology-neutral legislation cannot be legally certain, though, as generalizing for a spread of technologies does not pinpoint rulings for each. Legal certainty makes disputes simple to resolve because everything is well defined and understood. Without legal certainty, the rules become uncertain.

Soft law, general laws that cover wide domains of developing technologies, can fix the issue of technology neutrality versus legal certainty. By providing a working foundation to build further legislation off of, soft law takes the first step to an adaptive and fluid legislature, without having some technologies outside the domain of regulation. Specific councils for a certain technology can then develop more specific and fine-tuned regulation, effectively delegating from broad-scope legislation to fine-grained legal certainty, soundly establishing both.

Timeliness of regulation form the last two issues. Both premature and overdue liability regulation cause issues during technological development: premature regulation can stifle innovation and leave a field crippled and unexplored, where overdue regulation leaves the technology and its side effects unrestrained and potentially irreversible. Overdue regulation can also fail to compensate wronged individuals.

Further confounding this issue is the constant evolution of new technologies. Legislation cannot remain rigid over new technologies that it may not apply to. The shifting of development wears away at legal certainty, degrading the overall quality of the regulations. Without regular updates to maintain rulings, effectiveness decays.

Responsible research and innovation, an approach to evaluating innovations or products through the interactive participation of regulators, stakeholders, and the business itself can help alleviate the timing issue of legislation. Cyclically repeating this process maintains the accuracy of the rulings and keeps the different interest groups involved in the legislative process, further reinforcing those guidelines.

Leenes, Palmerini, Koops, Bertolini, Salvini, and Lucivero (2017) assert that regulation must be founded in the virtues and rights that each society holds. In this way, every society needs a specific set of legislation to fit their norms and values. Through these values, multi-level legislation, and cyclical upkeep of rulings, we can create smart, effective regulation. Now, I apply these techniques to the legal gaps robotization creates.

Legislatively Patching Robotization

Now I will propose solutions for the responsibility gaps artificial moral agency cannot fix by itself, working backwards through the previous list. I cannot go into the fine details of each issue and provide legal certainty alone. Instead I propose a few soft, umbrella regulations to provide a foundation to build more precise regulation.

Danaher (2016) considers a shift from retributive justice to reformative justice to relieve the retribution gap. Such a shift would suit machines much better: punishment has little effect on an unconscious machine. Additionally, those insistent on retribution could degrade trust in the legal system. Note that this shift need not apply to human agents as well.

Liability gaps can be spanned by a form of liability insurance falling on the manufacturer of the robots. The company would pay for this insurance to compensate those harmed by the manufacturer's product. Companies with higher rates of error would pay higher premiums for this kind of insurance, further incentivizing morally capable developments.

We can work towards minimizing moral responsibility gaps by having a fine — potentially covered by liability insurance — when an ethically-blind or morally-unprepared unit causes harm. With such a fee in place, companies have incentives to refine their artificial moral agents. Additionally, the fine would fall on those who did not develop moral machinery where an umbrella law required it. Mandatory physical or cloud-based recording devices, like those on airplanes, inside every moral agent could solve moral scapegoating. Such a device could record every detail and situation the machine encounters, determining how a machine came to cause harm. After analysis of the data, adjustments to future agents would reduce the chance of future damage.

Causal responsibility has not developed a significant gap, but proper analysis through the aforementioned recording system could further bring an end to “accidental damages.” True accidents will remain part of daily life. A technology-neutral definition of accident could further clear cases of causal ambiguity.

Thus, I conclude proposing preparations for legislative issues caused by further robotization. Shift from retributive blame to restorative reform can solve retribution gaps. Liability insurance can protect manufacturers and compensate victims of robotic error. A record of circumstances can resolve liability scapegoating and determine if a robot error properly classifies as an accident. Next, I move to a broader scale of issues caused by autonomous machines.

International Regulation

Much like our regulation of nuclear technology, robotics technology requires international regulation. AI weaponry will start a global AI arms race (Future of Life Institute, 2015). Based on the machine security argument above, I support an immediate international ban on developing AI weaponry, as AI weapons could easily lead to a global massacre in the wrong hands. Dissident groups capturing, controlling, and using such technology could have global-scale repercussions, potentially including a third world war (Musgrave & Roberts, 2015). Regardless of how ethically we program them and how morally right they are, reprogramming an autonomous weapon would free it for any purpose of the appropriator. Autonomous peacekeeping forces must not be armed or deployed by any government. Robo-cops could easily and efficiently disperse and end any protests or activism, which are the foundations for social change.

Through careful legislative change on the national and international levels, we can address the issues robotization causes in today’s world and prepare for future robotics development. It is crucial that we build an adaptive regulatory framework based on virtues, cyclical refinement, and interactivity between legislators, businesses, and stakeholders for future sustainable robotics regulation. I proposed some solutions to the responsibility issues of robotization, and discussed an international ban on AI weapons.

Conclusions

Morality for a machine consists of good action. We need ethical sensitivity in our machines to reduce the risk of harm to humans. Increased autonomy in a machine requires increased morality in said machine. However, we need not (and cannot) have perfection in our moral machines.

I discussed the methods of developing artificial moral agency. Top-down approaches develop fully operational ethics software before implementation, and include rule-based, virtue-based, and consequence-based systems. Bottom-up approaches learn moral functioning over time, and include neural networks, genetic algorithms, and scenario analysis systems. I proposed a union of a rule-based system reinforced by two neural networks: one for moral situations outside of the domain of the rule-based systems and one for explaining the decision making process of the backup system.

Artificial moral machines require embodiment, learning, empathy, and teleology towards good. I argued that because we cannot determine emotion or true intention in humans, we need not require those in our machines.

Responsibility gaps arise from the further development of autonomous machines. Causal responsibility gaps rise from robotic accidents. When an ethically-incapable machine morally errs, a moral responsibility gap appears. Liability deficit happens when a manufacturer no longer holds liability responsibility for a machine, and no one can compensate the victim of the accident. Retribution gaps happen when no party involved can be punished appropriately. I looked to legislation to cover these legal holes.

After discussing the difficulties of proper legislation — shifting technology, technology-neutrality versus legal certainty, and preemptive versus overdue regulation — I argued that we require further legislative action to bridge the gaps in our social and legal systems that robotization causes.

I conclude that while robotization causes problems, if we can prepare legally for those problems, we can integrate moral machines into our society without issue. We are very close to having artificial moral agents and the proper legislation to support them. With enough planning and care, we can assimilate moral machines into the world without destroying the systems we already have in place. Robotics development will continue to change the world, and with careful policy and regulation, we can surf the wave of robotics technology rather than be swallowed by it.

References

Allen, C., Varner, G., and Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence* 12(3). 251-261

Allen, C. (2011). The future of moral machines. *New York Times.* Retrieved from: https://opinionator.blogs.nytimes.com/2011/12/25/the-future-of-moral-machines/

Asimov, I. (1950). *I, Robot*. New York City: Doubleday.

Beavers, A. (2011). Moral machines and the threat of ethical nihilism. In Abney, K., Bekey, G. & Lin, P., *Robot ethics: The ethical and social implications of robotics* (333-344). Cambridge: MIT Press

Brynjolfsson, E., and McAfee, A. (2011). *Race against the machine.* Retrieved from: http://b1ca250e5ed661ccf2f1-da4c182123f5956a3d22aa43eb816232.r10.cf1.rackcdn.com/contentItem-5422867-40675649-ew37tmdujwhnj-or.pdf

Creighton, J. (2016, July 1). The evolution of AI: Can morality be programmed?. *Futurism*. Retrieved from https://futurism.com/the-evolution-of-ai-can-morality-be-programmed/

Danaher, J. (2016). Robots, law and the retribution gap. *Ethics and Information Technology, 18*(4), 299-309. doi:10.1007/s10676-016-9403-3 Retrieved from https://link.springer.com/content/pdf/10.1007%2Fs10676-016-9403-3.pdf

DeBaets, A. (2014). Can a robot pursue the good: Exploring artificial moral agency. *Journal of Evolution and technology*. 24(3) 76-86.

Dreyfus, H., and Dreyfus, S., (1992). What artificial experts can and cannot do. *AI & Society*. 6(1).

Frankena, W. (1973). *Ethics*. 2nd ed. Foundations of Philosophy Series. Englewood Cliffs, NJ: Prentice-Hall.

Fox, S. (2009). Evolving robots learn to lie to each other. *Popular Science*. Retrieved from: https://www.popsci.com/scitech/article/2009-08/evolving-robots-learn-lie-hide-resources-each-other

Future of Life Institute. (2015, July 28). Open letter on autonomous weapons. Retrieved from: https://futureoflife.org/open-letter-autonomous-weapons/

Gent, E. (2016, Oct 4). Device can read emotions by bouncing wireless signals off your body. *Live Science*. Retrieved from: https://www.livescience.com/56373-device-uses-wireless-signals-to-read-emotions.html

Goffman, E. (1956). *The presentation of self in everyday life*. New York: Random House.

Hellstrom, T. (2013). On the moral responsibility of military drones. *Ethics and Information Technology*, 15, 99–107.

Kuang, C. (2017, November 21). Can AI be taught to explain itself. *New York Times.* Retrieved from: https://www.nytimes.com/2017/11/21/magazine/can-ai-be-taught-to-explain-itself.html

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3), 175–183.

Mill, J. S. [1861] 1979. *Utilitarianism.* Indianapolis: Hackett Publishing Company.

Musgrave, Z., and Roberts, B., (2015, August 4). Humans, Not Robots, Are the Real Reason Artificial Intelligence Is Scary. *The Atlantic.* Retrieved from: https://www.theatlantic.com/technology/archive/2015/08/humans-not-robots-are-the-real-reason-artificial-intelligence-is-scary/400994/

Leenes, R., Palmerini, E., Koops, B.J., Bertolini, A., Salvini, P., and Lucivero, F., (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues. *Law, Innovation, and Technology, 9*(1). Retrieved from: https://www.tandfonline.com/doi/full/10.1080/17579961.2017.1304921

Papps, K., and Winkelmann, R., (1999). Unemployment and crime: New evidence for an old question. Retrieved from: https://www.nuffield.ox.ac.uk/users/papps/unemployment.pdf

Parkin, S. (2017, July 31). Teaching robots right from wrong. *The Economist*. Retrieved from: https://www.1843magazine.com/features/teaching-robots-right-from-wrong

Roy, S. (2017, April 12). How factory automation can reduce production costs. Retrieved from http://www.oemupdate.com/cover-story/how-factory-automation-can-reduce-production-costs/

Turkle, S. (2011). *Alone together: why we expect more from technology and less from each other.* New York: Basic Books.

Wallach, W., & Allen, C. (2009). *Moral machines: Teaching robots right from wrong*. Oxford: Oxford University Press.

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face; How individual face parts contribute to successful emotion recognition. *Plos One, 12*(5). doi:10.1371/journal.pone.0177239

Wilson, D., (2011). *Robopocalypse.* New York: Doubleday.

Yudkowsky, E., (2011). Complex value systems are required to realize valuable futures. In *Proceedings of AGI 2011*. Springer.