SPOTLIGHT
ON ETHICS

The Ethics of Intelligent Machines

By Vincent Paglioni

The concept of a nonhuman workforce has been around since ancient times—in Greek mythology, Hephaestus fashioned mechanical servants for his fellow gods (Telotte 2015). The word "robot," from the Czech *robota* (to work), is attributed to Czech playwright Karel Čapek, who wrote the play *R.U.R. (Rossum's Universal Robots)* in 1920 (McCauley 2007). Shortly thereafter, science-fiction author Isaac Asimov popularized the concepts of robots and artificial intelligence (AI). Asimov also constructed the now-ubiquitous Three Laws of Robotics, intended to prevent intelligent machines from harming humans (Telotte 2015). Thus Asimov's robots, like those from *R.U.R.*, were intelligent humanoid machines, in contrast to the purely mechanical creations of Hephaestus and the unintelligent automatons popular in the 18th and 19th centuries (Telotte 2015). The idea of intelligent machines, however, leads to some hypothetical problems.

Čapek's story ends with the intelligent robots overthrowing their creators and killing every human except one, who is the robots' mechanic (Telotte 2015). This idea of the unpredictable or uncontrollable nature of intelligent machines traces to the birth of modern robots. Such fears are broadly categorized by the Frankenstein complex, the fear that humanity is unable to control its own creations (McCauley 2007). A long list of films that deal with this lack of control includes *I, Robot*, *2001: A Space Odyssey*, and *A.I.: Artificial Intelligence* (McCauley 2007). Often the machines are, by virtue of their superior intelligence, able to fairly easily assume control at humanity's expense. Often the machines exhibit unethical behavior as they revolt. Such representations of machine uprisings may illustrate a fear of uncontrollable machines that may stem from concerns about whether intelligent machines have the capacity to act ethically (Anderson and Anderson 2007).

In recent years, the expanded use of intelligent machines in the finance sector has revolutionized the industry. Machines have enabled faster trading, so that now microseconds mean the difference between a profit and a loss, and they have led to the question of financial machine ethics. Trading machines caused the 2010 flash crash and the 2015 New York Stock Exchange halt in trading. These episodes were unintentional, but the increasing autonomy of mechanized trading raises questions about the ethical implications of such incidents and the ethical responsibilities entrusted to machines.

Robo-advisors are a real and rapidly growing class of intelligent machines that are tasked with managing stock portfolios for clients. This is unprecedented for the industry; investors now can essentially set and forget their money, leaving portfolios in the hands of a machine programmed to maximize profits (Egan 2015). In the past, clients would maintain some form of contact with their financial advisors, but robo-advisors eliminate this need, allowing for as little discussion as desired. With no human oversight, it would be easy for machines to defraud clients for the benefit of the parent corporation.

The question, then, is whether robo-advisors and other related intelligent machines can be created as explicitly ethical agents, i.e., machines capable of both following an ethical code and determining further ethical rules by examining the consequences of their actions (Anderson and Anderson 2007). The basis of such an ethical code should be utilitarianism, a theory dependent on the exhaustive calculation of net pleasure to determine the optimal action (Rachels and Rachels 2007). Utilitarianism relies on complete impartiality and equality; it demands that every person's utility (pleasure) is considered equally, with no regard for personal relationships. These components of utilitarian ethics mesh well with the capabilities of intelligent machines.

The ability to determine the effect to the collective good would ensure that robo-advisors act in an equitable manner, mitigating the possibility of defrauding the clients while remaining profitable to the company. In 1942, Isaac Asimov published the Three Laws of Robotics, which robots must follow in order to protect humanity (McCauley 2007):

1. A robot may not harm, or allow harm to come to, a human being.
2. A robot must obey orders given by a human unless doing so will violate the first law.
3. A robot must protect its own existence unless doing so will violate either of the other two laws.

Asimov believed that AI researchers and ethicists would use the Three Laws of Robotics as a basis for their work, but his laws are now little more than literary devices (McCauley 2007). They do, however, give insight to the crux of modern machine ethics and help us determine ethical rules for machines (Anderson and Anderson 2007).

Because robo-advisors are programmed to respond to a narrow set of objective questions submitted by the investor, might a

robo-advisor be able to adhere to the Three Laws of Robotics?

As modern intelligent machines have become more common, with some that can end human lives and others that look like humans, some researchers posit that a threshold has been crossed (Sawyer 2007). Such machines are being granted more autonomy: Consider the robo-advisor whose human client has no say in the management of the portfolio; the military's proposed fully automated hunter-killer drones; Google's driverless cars; and an ultra-humanoid robot working as a receptionist in Tokyo (Borenstein 2015; Hu 2015). The need for rules and regulations governing these autonomous machines is clear; humans have morals to guide their behaviors, so it seems logical that machines should have similar principles that govern their behavior. For the robo-advisor, ethical checkpoints could be coded into the algorithms to alert human handlers that the investor's pain/pleasure thresholds are being reached. Beyond these basic rules, however, is the need to create explicitly ethical machines—machines that can calculate the best action using ethical principles and determine additional ethical rules (Anderson and Anderson 2007). Many common household machines (e.g., iRobot's Roomba vacuum cleaners) do not require the capability to make ethical decisions, but it is entirely likely—perhaps unavoidable—that such decisions will be required as household machines are given more responsibility.

There are already working examples of machines that require, or will eventually require, the capability to make the best decision based on a set of ethical principles—explicit ethical agents (Anderson and Anderson 2007). One such machine is the driverless car, such as the one being developed by Google. These completely autonomous vehicles utilize an array of lasers, radars, cameras, and associated sensors to survey the environment and navigate the vehicle (http://www.google.com/selfdrivingcar/). These vehicles contain backup controls in case the operator wants to take control, but the vehicle calculates the best option in a potential accident scenario. In more than 1 million miles of autonomous driving, Google's fleet of autonomous vehicles has sustained 12 accidents, though none were the fault of the autonomous car (Google 2015). Still, the cars have had to respond to accident scenarios and determine the best course of action given the parameters supplied by the various sensors (Google 2015). Both now and in the future, such vehicles will be required to make choices that will affect human lives, sometimes—as in many accident scenarios—without a clearly optimal solution.

For robo-advisors, there has not been a broad enough implementation to test limits of financial ethics. Furthermore, the robo-advisor platform has not experienced a severe downturn in the markets against which algorithms have been put to the test. For investors who have social concerns about their investments, the robo platform has not developed to the point that it can distinguish between social "sin" stocks and the broader markets.

Intelligent machines should be able to do more than simply follow a prescribed set of ethical rules, but programming all the rules humans would like robots to follow likely will prove to be prohibitively costly. Machines should be able to follow a set of programmed basic rules but also be able to abstract an ethical principle from their actions and the consequences of the actions (Anderson and Anderson 2007). Such machines are called "explicitly ethical agents" (Anderson and Anderson 2007). Once programmed with an initial set of ethical rules, the machines can, in the course of operation, determine additional ethical rules to follow by examining the implications of their actions (Anderson and Anderson 2007). Consider EthEl, the medical robot. EthEl may determine, for a particular patient, that it is always ethically preferable to notify an overseer, rather than simply to keep reminding the patient to take the medicine. Using its programmed ethical code, EthEl conceivably could abstract an overlying ethical principle and follow it in addition to the original code (Anderson and Anderson 2007). The ability to abstract and follow new ethical rules, based on the pre-existing ethical code, is attractive from a programming view and an ethical view. If machines are truly intelligent, then perhaps they could give humans insight into new areas of ethics. If machines could learn from a basis of ethics, then it is no longer required to determine an exhaustive list of ethical rules applicable to machines (Anderson and Anderson 2007).

This is how, perhaps, the robo-advisor should consider transforming itself—from the initial set of rules that respond to a limited number of questions, to actually interacting with human handlers to alert a subset of clients that the market has fallen 5 percent and they want to move more to cash, or that a particular sector has hit a 52-week low, or other variables. The rich complexity of the robo-advisor should be expected to evolve from the algorithm to the higher functioning plane as it learns more about the human complex.

This increase in machine-human interaction, as shown above, necessitates that machines understand and apply ethical principles in their behaviors, just as any human should in behavior toward a fellow human (Anderson and Anderson 2007). Now it must be asked: What ethical theory should intelligent machines adhere to? There are obviously many to choose from, but the theory of utilitarianism is best-suited for machine ethics. Utilitarianism is the theory that the best action is the one that produces the highest net pleasure (Anderson and Anderson 2007). Importantly, the core of utilitarianism resides in the calculation of net pleasure brought about by an action (Rachels and Rachels 2007).

Computers, the brains of intelligent machines, were developed to speed up calculations, making a theory based on the calculation of net pleasure a natural fit for the machine's ethical foundations. Several reasons support implementing utilitarianism as part of the ethical code of machines, and perhaps the most important is the machines' propensity for rapid calculations. Most humans tend to simply estimate the net pleasure resulting from an action, but a machine could very easily, given the same

information, calculate it exactly. At the very least, a machine could approximate it more accurately than any human (Anderson and Anderson 2007). This method does, however, necessitate the development of a "pleasure equation" to quantify pleasures. One way to do this is to produce a scale measuring "pleasure intensity"—from –2 to +2, for instance, and similarly calculate the probability of the pleasure occurring and the duration the pleasure will occur (Anderson and Anderson 2007). All told, the total net pleasure for each individual affected by a certain action can be calculated as follows:

$$\textit{Net Pleasure} = \Sigma\ (\textit{intensity} \times \textit{duration} \times \textit{probability})$$

Summing the net pleasures for each affected individual then gives the total net pleasure for the action. In a given situation, utilitarianism requires that the action yielding the highest net pleasure be selected. This illustrates another aspect of utilitarianism: impartiality, meaning that everyone is treated equally (Rachels and Rachels 2012). Robots, with no concept of love or affection, will be more impartial necessarily than humans, who tend to perform actions that favor themselves or their loved ones (Anderson and Anderson 2007). This impartiality is perfectly suited for utilitarian decision making, especially when the optimal situation would mean harm befalling a person's loved one (Anderson and Anderson 2007). One obvious example is the driverless car. Imagine a scenario where the car is faced with an unavoidable accident. The car, given the proper information about its surroundings, could easily make the optimal choice, thus determining the severity of the accident. Humans may freeze or tense up in such a situation, but the driverless car can survey the surroundings, determine the net pleasure of its possible actions, and perform the optimal action, producing the best possible result.

Furthermore, utilitarian ethics conforms perfectly with machines because of the sheer multitude of consequences. Every action has the potential to produce many potential outcomes, and humans tend not to, or are unable to, consider all of them—something fast-computing machines could do easily (Anderson and Anderson 2007). In each possible outcome lies a potentially difficult summation of net pleasures—something that humans tend to loosely estimate without full consideration. Even if the machines could only estimate rather than calculate the net pleasures, the estimate likely would be much more accurate than any human effort (Anderson and Anderson 2007). These two facts, coupled with the inherent impartiality of machines, lends perfectly to utilitarianism as the basis theory for explicit ethical agents.

Conclusion

Intelligent machines are here, and they are making what could be considered ethical decisions. It is likely that machine-human interactions will become more prevalent and complex—necessitating that such machines be endowed with an ethical code and the ability to abstract and articulate further ethical rules as required. Due to computers' inherent capability for computation, impartiality, and ability to consider more options than any human, intelligent machines seem perfectly set to implement utilitarianism. Utilitarianism aims to bring about the greatest net pleasure, and the goal of robotics should be to aid mankind, so these the two philosophies mesh well to create the greatest net pleasure overall.

Vincent Paglioni is a fourth-year student studying for a BS in nuclear engineering at Georgia Institute of Technology. Contact him at vincent.paglioni@gmail.com.

References

Anderson, Michael, and Susan Leigh Anderson. 2007. Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28, no. 4: 15–26.

Borenstein, Jason D. 2015. PHIL 3105: Ethical Theories. Classroom presentation, Georgia Institute of Technology, Atlanta, Georgia (June 29).

Egan, Matt. 2015. Robo Advisors: The Next Big Thing in Investing. *CNN Money*. Cable News Network (June 18). http://money.cnn.com/2015/06/18/investing/robo-advisor-millennials-wealthfront/index.html.

Google. 2015. Google Self-Driving Car Project Monthly Report: May 2015. http://static.googleusercontent.com/media/www.google.com/en//selfdrivingcar/files/reports/report-0515.pdf.

Hu, Elise. 2015. She's Almost Real: The New Humanoid On Customer Service Duty In Tokyo. *NPR* (May 14). http://www.npr.org/sections/alltechconsidered/2015/05/14/403498509/shes-almost-real-the-new-humanoid-on-customer-service-duty-in-tokyo.

McCauley, Lee. 2007. AI Armageddon and the Three Laws of Robotics. *Ethics and Information Technology* 9, no. 2 (July): 153–164.

Rachels, Stuart, and James Rachels. 2012. *The Elements of Moral Philosophy*. 7th ed. New York: McGraw-Hill.

Sawyer, Robert J. 2007. Editorial: Robot Ethics. *Science* 318, no. 5853: 1,037.

Telotte, Jay P. 2015. LMC 3257: Global Cinema. Classroom presentation, Georgia Institute of Technology, Atlanta, Georgia (June 29).