Amitai Etzioni and Oren Etzioni

Designing AI Systems that Obey Our Laws and Values

Operational AI systems (e.g., self-driving cars) need to obey both the law of the land, and our value system. We propose AI oversight systems ("AI Guardians") as an approach to addressing this challenge, and to respond to the potential risks associated with increasingly autonomous AI systems.[1] These AI oversight systems serve to verify that operational ones did not stray unduly from the guidelines of their programmers and to bring them back in compliance if they do stray. The introduction of such second-order, oversight systems is not meant to suggest strict, powerful, or rigid (from here on 'strong') controls. Operations systems need a great degree of latitude in order to follow the lessons of their learning from additional data mining and experience and to be able to render at least semi-autonomous decisions (more about this below). However, all operational systems need some boundaries, both in order to not violate the law and to adhere to ethical norms. Developing such oversight systems, AI Guardians, is a major new mission for the AI community.

All societies throughout history—in the pre-AI world and the current parts of the world that are not digital—have had oversight systems. Workers have supervisors; businesses have accountants; school teachers have principals. That is, all these systems have hierarchies in the sense that the first line operators are subject to oversight by a second layer and are expected to respond to corrective signals from the overseers. (These, in turn, are expected to take into account suggestions or even demands by the first line to change their modes of oversight). John Perry Barlow, in his famous Declaration of the Independence of Cyberspace in 1996 described the burgeoning online world as one that would be governed by a social contract formed among its users.[2]

---

[1] See Thomas G. Dietterich, Eric J. Horvitz *Communications of the ACM (Oct. 2015), "Rise of Concerns about AI: Reflections and Directions"* for an in depth discussion of the various risks.
[2] https://www.eff.org/cyberspace-independence

Terra incognita

AI systems not only need some kind of oversight, but this oversight must be provided—at least in part— not by mortals, but by a new kind of AI system, the oversight ones. AI needs to be guided by AI.[3]

One reason is that AI operational systems are *learning systems*. These systems do not stop collecting data once they are launched; instead, continued data mining and experience are used to improve their performance. These AI systems may hence stray considerably from the guidelines their programmers initially gave them. But no mortal can monitor these changes, let alone in real time, and determine whether they are legal and ethical.

Second, AI systems are becoming highly *opaque*, "black boxes" to human beings. Jenna Burrell from the School of Information at UC-Berkeley distinguishes three ways that algorithms become opaque: 1) Intentional opacity, for example with proprietary algorithms that a government or corporation wants to keep secret. 2) Technical illiteracy, where the complexity and function of algorithms is beyond the public's comprehension. And 3) Scale of application, where either "machine learning" and/or the number of different programmers involved renders an algorithm opaque even to the programmers.[4]

Finally, AI guided systems have considerable *autonomy* in the sense that they make numerous choices "on their own."[5] That is, these instruments, using complex algorithms, respond to environmental inputs independently.[6] They may even act in defiance of the guidelines the original programmers installed. A simple example is automatic emergency

---

[3] See Weld & Etzioni (AAAI, 1994) "The First Law of Robotics (a call to arms)" for an early attempt to formalize a solution to this problem.
[4] Burrell, Jenna. "How the machine 'thinks': Understanding opacity in machine learning algorithms." *Big Data & Society* 3, no. 1 (2016).
[5] Viktor Mayer-Schönberger & Kenneth Cukier, *Big Data* (2014), 16-17.
[6] *New algorithm lets autonomous robots divvy up assembly tasks on the fly*, SCIENCE DAILY, (May 27, 2015), http://www.sciencedaily.com/releases/2015/05/150527142100.htm.

braking systems,[7] which stop cars without human input in response to perceived dangers.[8] Consumers complain of many false alarms, sudden stops that are dangerous to other cars,[9] and that these brakes force cars to proceed in a straight line even if the driver tries to steer them elsewhere.

For all these reasons, AI oversight systems are needed. We call them AI Guardians. A simple dictionary definition of a guardian is: "a person who guards, protects, or preserves.[10]" This definition captures well the thesis that oversight systems need not be strong because this would inhibit the innovative and creative development of operational AI systems—but cannot be avoided. Indeed, a major mission for AI is to develop in the near future such AI oversight systems. (We discuss below whose duty is to develop these oversight systems and to whom they are to report their findings and whose values they are to heed).

Different kinds of AI Guardians

Interrogators: After a series of crashes of drones manufactured by one corporation, another corporation that purchased several hundred drones is likely to try to determine the cause of the crashes. Were they intentional (e.g. caused by workers opposed to the use of drones)? Unwitting flaws in the design of the particular brand of drones? Flaws in the AI operational system that serves as the drone's 'brain'? For reasons already discussed, no human agent is able to provide a definitive answer to these questions. One would need to design and employ an interrogator AI system to answer the said questions.

---

[7] Chris Kapnan, *Auto-braking: a quantum leap for road safety*, THE TELEGRAPH, (August 14, 2012), http://www.telegraph.co.uk/motoring/road-safety/9429746/Auto-braking-a-quantum-leap-for-road-safety.html.
[8] Mark Phelan, *Automatic braking coming, but not all systems are equal*, DETROIT FREE PRESS, (January 1, 2016), Available at http://www.freep.com/story/money/cars/mark-phelan/2016/01/01/automatic-braking-safety-pedestrian-detection-nhtsa-iihs/78029322/.
[9] Eric Limer, *Automatic Brakes Are Stopping for No Good Reason*, POPULAR MECHANICS, (June 19, 2015), www.popularmechanics.com/cars/a16103/automatic-brakes-are-triggering-for-no-good-reason/.
[10] http://www.dictionary.com/browse/guardian?s=t

In recent years, several incidents show the need for such interrogation. In 2015, a team of researchers from Carnegie Mellon University and the International Computer Science Institute found that Google was more likely to display ads for high-paying executive jobs to users that its algorithm believed to be men than to women.[11] Google stated that there was no intentional discrimination but that the effect was due to advertisers' preferences.[12]

In 2014, Facebook conducted a study unbeknownst to its users wherein its algorithms manipulated users' posts to remove "emotional content" in order to gauge reactions from the posters' friends.[13] Facebook later apologized for not informing its users about the experiment. Twitter recently deleted 125,000 accounts, stating that these included only accounts that were linked to the Islamic State.

If a committee of the board of these corporations or an outside group sought to verify these various claims—they would need an AI monitoring system.

Auditor: Wendell Wallach, a scholar at Yale's Interdisciplinary Center for Bioethics points out that "in hospitals, APACHE medical systems help determine the best treatments for patients in intensive care units—often those who are at the edge of death. Wallach points out that, though the doctor may seem to have autonomy, it could be very difficult in certain situations to go against the machine—particularly in a litigious society."[14] Hospitals are sure to seek audits of such decisions and they cannot do so without an AI auditing system.

Monitor: Because self-driving cars are programmed to learn and change, they need a particular kind of AI Guardian program –an AI Monitor – to come along for the ride to ensure

[11] https://www.technologyreview.com/s/539021/probing-the-dark-side-of-googles-ad-targeting-system/
[12] https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study
[13] https://www.theguardian.com/technology/2014/jun/29/facebook-users-emotions-news-feeds
[14] http://qz.com/653575/can-we-trust-robots-to-make-moral-decisions/

the autonomous car's learning does not lead it to violate the law, for example learning from the fact that old fashioned cars violate the speed limit and emulating this behavior.

Enforcer: In rare situations, an AI Guardian may help enforce a regulation or law. For instance, if the computers of a military contractor are repeated hacked, an AI enforcer may alert the contractor that it needs to shore up its cyber defenses. If such alerts are ignored, the AI enforcer task will be to alert the 'clients' of the contractor or to suspend its clearance.

Ethics bots: AI operational systems must not only abide by the law, but also heed the moral norms of society. Thus driverless cars need to be told whether they should drive at whatever speed the law allows, or in ways that conserve fuel to help protect the environment, or to stay in the slower lanes if children are in the car. And—if they should wake up a passenger in the back seat if they "see" an accident.

Several ideas have been suggested as to where AI systems may get their ethical bearings. In a previous publication, we showed that asking each user of these instruments to input his or her ethical preferences is impractical, and that drawing on what the community holds as ethical is equally problematic. We suggested that instead one might draw on ethics bots.[15]

An ethics bot is an AI program that analyzes many thousands of items of information—not only information publicly available on the Internet but also information gleaned from a person's own computers about the acts of a particular individual that reveal that person's *moral* preferences. And then uses these to guide the AI operational systems (for instruments used by individuals, such as the driverless cars).

[15] Amitai Etzioni and Oren Etzioni. 2016. "AI Assisted Ethics." *Ethics and Information Technology,* Vol 18, No. 2 (pp. 149–156). http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2781702

Essentially, what ethics bots do for moral choices is similar to what AI programs do when they ferret out consumer preferences and target advertising accordingly.[16] In this case, though, the bots are used to guide instruments that are owned and operated by the person, in line with their values—rather than by some marketing company or (political campaign). For instance, such an ethics bot may instruct a person's financial program to invest only in socially responsible corporations, and in particular green ones, and make an annual donation to the Sierra Club, based on the bot's reading of the person's past behavior.

In short, there is no reason for the digital world to become nearly as hierarchical as the non-digital one. However, the growing AI realm is overdue for some level of guidance to ensure that AI operational systems will act legally and observe the moral values of those who own and operate them.

It is not necessarily the case that AI guardians are more intelligent than the systems they oversee. Rather, the guardians need to be sufficiently capable and intelligent that they are not outwitted or short-circuited by the systems they are overseeing. Consider, for example, a electrical circuit breaker in a home: it is far less sophisticated than the full electrical system (and associated appliances) it is protecting but it is quite reliable, and can be "tripped" by a person in an emergency.

---

[16] For example, Nielsen has developed a marketing system for targeting very specific demographics with financial and investment products based on age, affluence, the presence of children in the home, and certain purchasing habits. These include such specific target consumer groups as "Y2-54: City Strivers" and "F4-56: Economizers." "Nielsen P$YCLE Lifestage Groups," Accessed December 17, 2015, Available at https://www.claritas.com/MyBestSegments/Default.jsp?ID=8010&pageName=Learn%2BMore&menuOption=learnmore.
Ted Cruz' campaign in Iowa relied on psychological profiles to determine the best ways to canvass individual voters in the state. Tom Hamburger, "Cruz campaign credits psychological data and analytics for its rising success," *The Washington Post*, December 13, 2015, Available at https://www.washingtonpost.com/politics/cruz-campaign-credits-psychological-data-and-analytics-for-its-rising-success/2015/12/13/4cb0baf8-9dc5-11e5-bce4-708fe33e3288_story.html.

AI researchers can work towards this vision in at least three ways. First, they can attempt to formalize our laws and values following an approach akin to that outlined in the work on formalizing the notion of "harm".[17] Second, researchers can build labeled data sets identifying ethical and legal conundrums labeled by desired outcomes, and provide these as grist for machine learning algorithms. Finally, researchers can build "AI operating systems" that facilitate off switches as in the work on "safely interruptible agents" in reinforcement learning.[18] Our main point is that we need to put AI guardian on the research agenda for the field.

Who will guard the AI Guardians?

There are two parts to this question. One concerns who will decide which AI oversight systems will be mobilized to keep in check the operational ones. Some oversight systems will be introduced by the programmers of the software involved at the behest of the owners and users of the particular technologies. For example, those who manufacture driverless cars and those who use them will seek to ensure that their cars will not speed ever more. This is a concern as the cars' operational systems—which, to reiterate, are learning systems—will note that many traditional cars on the road violate the speed limits. Other AI oversight systems will be employed by courts and law enforcement authorities. For instance, in order to determine who or what is liable for accidents, and whether or not there was intent.

Ethics bots are a unique AI Guardian from this viewpoint. They are to heed the values of the user, not the owner, programmer, or those promoted by the government. This point calls for some explanation. Communities have two kinds of social and moral values. One kind includes values the community holds, which are of particular importance and hence their implementation cannot be left to individual choice; heeding them is hence enforced by coercive means, by the law. These values include a ban on murder, rape, theft and so on. In the AI world,

[17] Weld and Etzioni (AAAI, 1994): The First Law of Robotics (a call to arms).
[18] https://intelligence.org/files/Interruptibility.pdf

heeding these is the subject of a variety of AI Guardians, outlined above. The second kind of values concern moral choices the community hold it can leave to each person to decide whether or not to follow. These values include whether or not to donate an organ, give to charity, volunteer, and so on. These are implemented in the AI world by ethics bots.

The question of who will guard the guardians rises. Humans should have the ultimate say about the roles and actions of both the AI operational and AI oversight systems; indeed all these systems should have an on and off switch. None of them should be completely autonomous. Ultimately, however smart a technology may become, it is still a tool to serve human purposes. Given that those who build and employ these technologies are to be held responsible for their programing and use, these same people should serve as the ultimate authority over the design and operation and oversight of AI.

Amitai Etzioni is a University Professor at The George Washington University. Oren Etzioni is CEO of the Allen Institute for Artificial Intelligence.