The Judges' Book

Volume 3 Article 17

2019

Technology Law: Artificial Intelligence: Trust and Distrust

Robin C. Feldman

Technology Law:
Artificial Intelligence: Trust and Distrust[1]

Robin C. Feldman

Artificial intelligence (AI) is percolating through modern society. In the automobile industry, AI systems assist drivers with steering, changing lanes, and parking. Early AI projects in the criminal-justice system predict where crime is likely to occur for the purpose of targeting policing. Smart glasses tailored to business applications are emerging into the marketplace. Eventually, these glasses will use machine learning to identify objects and voices, prompting the wearer to take certain actions or setting out a range of possible actions. Banking and insurance firms use AI to advise customers on financial services, assess consumer risk, and monitor for fraud. Employers use AI systems in hiring decisions. And in the healthcare field, invasive brain interfaces have demonstrated the ability for thought control of complex robotic limbs and virtual agents.

As AI becomes a ubiquitous part of our everyday life, a key aspect will be the way in which society—and by extension, the legal system—manages both the integration of these systems and society's expectations. Society will have to learn to trust the capacity of AI systems sufficiently so that it can soar to new heights, without succumbing to the "irrational exuberance"[2] that can send society crashing to the ground when AI fails to live up to unreasonable expectations. And society must learn to tolerate the ambiguity that lies between these two extremes.

The State of AI

AI refers to any artificial agent that ingests data and responds to that data by completing a task. The AI we have today is "narrow AI": AI that can complete only one or a handful of pre-specified tasks. We are nowhere near "Artificial General Intelligence—defined by its potentially unattainable ability to complete any task a human can, short

[1] Summarized and excerpted from Robin C. Feldman, *Artificial Intelligence: The Importance of Trust and Distrust*, 21 GREEN BAG 2D 201 (2018).

[2] Alan Greenspan, *Francis Boyer Lecture of the American Enterprise Institute for Public Policy Research: The Challenge of Central Banking in a Democratic Society* (Dec. 5, 1996).

of consciousness—let alone the type of AI that resembles a conscious mind.

Still, advancement in AI has moved at an extraordinary pace. Deep learning, which is the basis for the entire field of AI, became practical in 2006, and, since then, the field has moved in leaps and bounds, analogous to what would be many lifetimes in other industries. Consider perhaps the highest profile modern network, Google's AlphaGo, for which there have been at least four different versions since 2015. The difference between the two latest versions of AlphaGo would be analogous to the difference between the first rudimentary touchscreen phone—the IBM Simon from 25 years ago—and this year's new iPad Pro. In the AI field, however, that advancement happened in under two years.

The potential for truly sentient AI that can make decisions and operate on its own, however, remains in the minds of science fiction writers. For now, and for the foreseeable future, human augmentation systems will be the norm, and the optimal configuration will be a melding of human and machine capability. An example can be found in labor organization. Siemens is currently developing a factory in which jobs are assigned to human workers by AI that knows a worker's skill. As a start, the AI will assign jobs that require human dexterity to humans while assigning jobs that can be done by robots to robots. As robotic dexterity improves across time, the "boss AI" can assign jobs to humans on the basis of other skills that robots lack, such as language and reasoning. As one of the researchers noted, one would not want to reduce humans to mere tools in any system because then "we would just use an expensive human as an imprecise robot. When it comes to creativity and complex, intelligent tasks, this is where humans are superior." The goal is to "build systems that combine strengths from both sides."[3]

In short, for the foreseeable future, the best approaches are likely to be systems that can augment human capacity, rather than systems that replace human beings and operate entirely on their own. For those who are movie buffs, think Iron Man, in which a weaponized suit enhances the protagonist's capacities, as opposed to the Terminator, in which a machine-like cyborg does everything by itself. And even that imagery may be optimistic. As one expert commented to me, we can't have anything remotely like Iron Man because machines are just plain dumb. We still have to teach them what a stop sign is, and we are light years from a machine's ability to think on its own.

[3] Sean Captain, *This AI Factory Boss Tells Robots and Humans How to Work Together*, FAST CO. (Aug. 7, 2017) (quoting Florian Michahelles).

How Trust and Distrust Can Enhance Vulnerabilities

On the simplest level, people will have to be coaxed into using these newfangled devices. This is not just a matter of encouraging those who are older than 40 to use social media. Absent widespread usage, the full potential of AI systems may be limited.

Consider the potential for true driverless cars, not the driver-assisted versions that exist today, but cars that operate without any driver at all. To achieve its maximum potential, driverless cars will be linked into networks with other driverless cars on the road. Your car will not just slow down when it senses that the car in front of you has slowed down; your car could react when the network tells it that a car 10 blocks ahead has altered its speed or trajectory. With a networked system of this sort, particularly one that can react faster than humans, cars will need less space between them, and traffic flows can be maximized so that riders spend less time on the road and consume less fuel.

Imagine the difficulties that arise if every now and then, we mix in a human driver. The safety and efficiency calculations become much more complex and challenging as we increase the level of uncertainly—both the uncertainty of whether a car down the road is human driven as well as the uncertainty of what the human driver will choose to do.[4] In fact, in the current tests of driverless car systems, some of the greatest difficulties flow from interacting with human drivers who are puzzlingly irrational. The point is simply that some of the power of AI systems depends not just on whether humans can be coaxed into using them at all but also whether that use is widespread, even ubiquitous.

Trust has other facets as well. From a different perspective, both government and individuals in society will need to have confidence in the actions and choices made by AI technologies. If we want ordinary citizens to have faith in the credibility of AI, there must be methods of analyzing and validating the choices made—trust but verify, as the old saying goes.

The entire issue of verification is complicated by the black box nature of certain AI systems. When decisions are being made that result in sending criminals to jail or choosing between killing the driver of an autonomous vehicle and a crowd of six, how do we develop the pathways for interrogating the technology to society's satisfaction?

[4] *See* Matt Richtel & Conor Dougherty, *Google's Driverless Cars Run into Problems: Cars with Drivers*, N.Y. TIMES (Sept. 1, 2015).

And then, how do we translate that verification into language that will inspire confidence among all citizens?

Both of these tasks will require a level of openness and candor that are not necessarily familiar to either industry or government players. In particular, a company's first instinct is unlikely to encompass throwing open the doors to its technology, particularly if competitors are peering into the open doorway. Nevertheless, one cannot expect citizens to gain trust in AI simply because we say soothingly, "Don't worry. We've got this covered." And the results of lack of trust can be far-reaching. What happens if all citizens, or even only certain groups of citizens, believe they cannot trust any information they are receiving on any level? In that circumstance, the breakdown of trust can be more serious than the disarray that can develop when individuals opt out of a linked network.

While trust is essential, overconfidence is detrimental. Overconfidence in AI can lead to unrealistic expectations of AI's capacity, accuracy, and security. Human judgment and interpretation remain crucial for responding to and recovering from the types of security incursions that AI technologies will face. Just as one would not build a fence around a power plant and consider the plant to be secure, one cannot simply set up cybersecurity perimeters and consider the job done. The strength of any networked system lies in its resilience after a vulnerability has been exploited or has caused harm.

AI systems will be no exception. Consider automobiles. When networked, driverless cars become the norm, and vulnerabilities increase exponentially. Any point throughout the vast network of cars becomes a potential door for malicious entry, and the damage may be far greater, given the extent of the network. A hacker only need find one point of vulnerability throughout all of the cars and their car systems in order to make every car in the network run off the road.

AI technologies are no more impregnable than any other technology. In particular, machine learning technologies are dependent upon their input data. When attackers corrupt that data, the system will continue to think it is operating properly. One study, albeit a limited one, managed to confuse automated systems about the nature of stop signs by placing unobtrusive stickers on the signs.[5] Another study, of medical devices, showed that heart pacemakers can be hacked.[6]

[5] *See* Jonathan M. Gitlin, *Hacking Street Signs with Stickers Could Confuse Self-Driving Cars*, ARSTECHNICA (Sept. 1, 2017).

[6] *See* Natt Garun, *Almost Half a Million Pacemakers Need a Firmware Update to Avoid Getting Hacked*, THE VERGE (Aug. 30, 2017).

The mortal consequences of these vulnerabilities make human supervision imperative. If an AI boss system is compromised and assigns English speakers to jobs that require Spanish proficiency, a human can easily detect this type of mistake. Even more subtle attacks, like assigning the wrong specialist to a cybersecurity project, can be detected relatively easily by a human familiar with the project and its personnel.

Even bread-and-butter data analysis is likely to work best with a combination of human and machine contributions. Large streams of data are impossible for humans to inspect by hand. Nevertheless, humans are far better than machines at playing detective, that is, noticing something that just does not seem right or finding an indication that points to a malware incursion and applying the creativity to figure out what is going on. Thus, AI may be best for sorting network traffic into smaller, human-manageable groups of information that the more creative human counterparts can oversee.

Perhaps the greatest potential risk of widespread adoption of AI is fear-inducing disruption. Imagine an attacker who changes the manufacturing instructions for a single bottle of a medication, or a hacker who alters the pattern for one person's pacemaker. Although most of the medications or medical devices are perfectly fine, widespread fear could lead to great harm if patients refuse to take their medication, decline to have pacemakers installed, or demand to have them removed. Panic, underlying anxiety, and erosion of trust can be widespread and culturally significant.

In this way, trust and distrust can wrap back around each other to maximize the risk of chaos and societal disruption. Imagine a time in which each person has an implanted health device, call it a health regulator. The device contains that person's health information, monitors various bodily functions, and can even direct other implantable devices such as pacemakers or automated medicine-dispensing mechanisms. Say a recent immigrant from North Korea is seriously injured and comes to an urgent care center. The patient's health regulator alerts the medical team to the need for a blood transfusion and indicates that the patient's blood type is AB negative. The team becomes suspicious because the blood type AB negative is almost nonexistent in the Korean population. Further investigation reveals that the person's health data was likely corrupted intentionally and that the problem most likely extends to other immigrants from North Korea.

The fallout could be extensive. Patients from North Korea might refuse to receive medical treatment. Those fears could cascade throughout immigrant populations, or throughout patient populations in general. The public in general or smaller groups in particular could mistakenly believe that the corruption extends beyond the North Korean immigrant population. People might fear that this corruption could be the beginning of larger incursions, either by this attacker or by others. Meanwhile, the entire health system, accustomed to relying on the efficiency of its health regulators, would be thrown into disarray as medical professionals must decide how to treat patients and make medical decisions without that input, not to mention what information and devices remain reliable.

The potential social implications also are profound. Disruption of the healthcare system connected to a recent immigrant population creates the potential for backlash against immigrant populations in the United States and abroad. Distrust of information in general could cascade to make various populations, particularly vulnerable populations such as new immigrants, unwilling to trust any information from the government, whether it is about the recent incursion in particular or health care in general. Such an outcome could lead immigrant populations to look for other, perhaps less reliable, sources of information. In short, a small and limited incursion could have extensive and profound effects on social cohesion and societal resources.

The Problems with Reactive Adaptation

In 1982, seven people died after taking Tylenol capsules adulterated with cyanide, an event that led to changes in medical packaging and to the creation of anti-tampering laws.[7] One might think of this history as a fine analogy—a blueprint that the legal system may use in adapting modern legal systems to manage issues created by AI. The government's reaction in the Tylenol case, however, was no more than a reaction, and reactive jurisprudence is seriously limited.

The problem lies beyond the fact that when legal systems adapt in reaction, damage has already occurred. Nor is the problem simply that one may not think clearly in the middle of a crisis. The real problem is that by the time one chooses to react, the choices may be limited. Within the social compact, we relinquish certain liberties related to the ends for which we have united, but it behooves us to decide which liberties to relinquish and which to nurture at a time when we still have sufficient choices available. Nowhere is this maxim more critical than at the dawn of a scientific revolution.

[7] *See* Ronald Reagan, *Statement on Signing the Federal Anti-Tampering Act* (Oct. 14, 1983).

Science is not immune to the dictates of the legal system. Rather, science and law exist in a symbiotic relationship, with each having the ability to inform or obstruct the other. Science creates pathways that drive legal regimes, because law cannot dictate what science cannot accomplish. In turn, law affects the unfolding of scientific development and shapes the expectations of individual citizens. When a car driven by a 16-year-old hits my car on the road, I expect the driver (or the driver's insurance company) to pay for the damage. I generally don't expect remuneration from the local authorities, who made the bad judgment to grant a license to this 16-year-old, or from the driver's parents, whose loose parenting styles might have influenced the level of driving care.

These byways, into which we channel both human expectation and scientific development, are best carved with thoughtful intention. The key will be to ensure that as these technologies permeate society, we design legal systems that embody appropriate levels of both trust and distrust.

Within this context, one of the challenges to openness and access could be a rush to secure intellectual property rights in AI. Trade secrets are, quite simply, secret. Patents can also fail to provide complete transparency; particularly in fields related to artificial invention such as software, current doctrines require only that patents disclose the outcomes of the invention, not how to get there.[8] Some other forms of invention rights might be needed.

In addition, AI systems should be subject to review entirely outside the system itself by industry bodies or public bodies. As an average citizen, I may never understand how a biologic interchangeable is being produced, at least not enough to trust that the drug is safe. Nevertheless, I might trust the FDA. This form of institutionalized outside review, whether by private or public entities, will be essential for adequate trust and distrust.

Regardless of the final routes chosen, the point is simply that society has an opportunity to craft legal regimes for AI on a broad scale. We are at the dawn of an era. Rather than stumbling blindly, we should move with care and intention.

[8] *See* ROBIN FELDMAN, RETHINKING PATENT LAW 104–27 (2012).

***