ARTIFICIAL INTELLIGENCE

THE IMPORTANCE OF TRUST AND DISTRUST

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 health care 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"[1]

Robin Feldman is a Visiting Professor at UCLA Law and the Harry & Lillian Hastings Professor of Law & Director of the Institute for Innovation Law at the University of California Hastings College of the Law. Copyright 2018 Robin C. Feldman.

[1] Alan Greenspan, Chairman, Fed. Reserve Bd., Remarks at the Annual Dinner and Francis Boyer Lecture of the American Enterprise Institute for Public Policy Research: The Challenge of Central Banking in a Democratic Society (Dec. 5, 1996).

that can send society crashing to the ground when AI fails to live up to people's blind expectations. And society must learn to tolerate the ambiguity that lies between these two extremes.

I.
THE STATE OF AI

What is referred to by the term "AI" is typically partitioned into two distinct categories corresponding to different levels of the characteristic we intuitively understand as intelligence.[2] The first, called "weak AI," refers to any artificial agent that ingests data and responds to that data by completing a task. No matter how complicated the input data or the task, weak AI at best does no more than create the appearance of having a "mind" or "consciousness." The second sort, "strong AI," can do everything weak AI can do, but is said to possess "a mind" and be truly capable of thought.

All current AI is weak, and there is a very serious debate in the fields of computer science and philosophy over whether that can ever change. Weak AI is further subdivided into "narrow AI" and "Artificial General Intelligence." This distinction is far more practical than the previous one. Narrow AI is defined as AI that can complete only one or a handful of prespecified tasks; Artificial General Intelligence is defined by its potentially unattainable ability to complete any task a human can, short of consciousness. The AI we have today is weak and narrow. We are nowhere near Artificial General Intelligence, let alone the type of AI that resembles a conscious mind.

One implementation, currently successful and in vogue, is the oft referred to "deep learning." A full primer on the subject is beyond the scope of this article, but learning is an umbrella term that refers to a particular kind of mathematical model for analyzing data. These models, called neural networks by analogy (and only by analogy) to a conception of how the human brain operates, analyze data by applying a series of mathematical transformations. Each transformation in the series is referred to as a "layer" of the neural network and deep learning refers to the field which con-

[2] *See generally* John Searle, *Minds, Brains, and Programs*, 3 BEHAVIORAL AND BRAIN SCIENCES 417 (1980).

cerns itself with neural networks having "many" layers. The iterative process involved in the many layers is referred to as "training" the neural network and the data is called the "training data." Having large amounts of good data is essential for any deep learning project.

Advancement in AI has moved at an extraordinary pace. Deep learning, which is the basis for the entire field of AI, was not practical until a 2006 paper[3] opened the door for quickly training neural nets. Since then, the field has moved in leaps and bounds, analogous to what would be many lifetimes in other industries. In fact, the basis for most modern neural nets, which rely on a class of models known as generative adversarial models, only emerged in 2014.[4]

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 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.

For many, the notion of AI brings to mind an eerie computer voice announcing annihilation of those humans in the vicinity, if not the elimination of the entire human race, or at least some form of takeover in which humans are reduced to slaves serving the ever more powerful machine masters. However, the potential for truly sentient AI that can make decisions and operate on its own 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. Consider chess, the basic bellweather for AI development. The 2005 Playchess.com tournament included teams of humans, teams of computers, and mixed teams. The tournament winner was a group of amateur chess players using three powerful computers.

A more practical example of human augmentation systems can be found in labor organization. Siemens is currently working on a factory in which jobs are assigned to human workers by AI that knows a worker's skill. As a

[3] *See* Geoffrey E. Hinton, Simon Osindero & Yee-Whye Teh, *A Fast Learning Algorithm for Deep Belief Nets*, 18 NEURAL COMPUTATION 1527 (2006).

[4] *See* Ian J. Goodfellow et al., *Generative Adversarial Nets*, NEURAL INFO. PROCESSING SYS. PROC. (2014), papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

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 on the Siemens project 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."[5] The goal is to "build systems that combine strengths from both sides."[6]

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.[7] 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. One can think of this human-AI interface, not just as a screen sending information to a human, but as a human-machine fusion, in which each can enhance the other or a form of augmented intelligence.[8] And even that imagery may be optimistic. As one expert commented to me, we can't have anything remotely like Ironman because machines are just plain dumb. We still have to teach them what a

[5] *See* Sean Captain, *This AI Factory Boss Tells Robots & Humans How to Work Together,* FAST COMPANY (August 7, 2017) (citing Florian Michahelles who heads the Siemens Web of Things research group in Berkeley, California), www.fastcompany.com/3067414/robo-foremen-could-direct-human-and-robot-factory-workers-alike.

[6] *See id.*

[7] *See* Ryan Calo, *Robotics and the Lessons of Cyberlaw*, 103 CALIF. L. REV. 513, 517, 539 (2015) (explaining that "robotics will continue to evolve, but mostly in ways that solve known technical challenges and reduce costs" and noting that "[l]ittle is gained, and much is arguably lost, by pretending contemporary robots exhibit anything like intent").

[8] *See generally* DOUG ENGLEBART, STANFORD RESEARCH INSTITUTE, AUGMENTED HUMAN INTELLECT STUDY (1962) (report prepared for the Air Force Office of Scientific Research) (coining and defining the term), web.stanford.edu/dept/SUL/library/extra4/sloan/mousesite/EngelbartPapers/B5_F18_ConceptFrameworkPt1.html; *see also* Editorial, *Anticipating AI*, 532 NATURE 413 (Apr. 28, 2016) (noting that "advances in robot vision and hearing, combined with AI, are allowing robots to better perceive their environments," which could lead to an "explosion of intelligent robot applications – including those in which intelligent robot applications work closely with humans), www.nature.com/polopoly_fs/1.19825!/menu/main/topColumns/topLeftColumn/pdf/532413a.pdf.

stop sign is, and we are light years from a machine's ability to think on its own.

II.
HOW TRUST AND DISTRUST CAN ENHANCE VULNERABILITIES

Like a multifaceted jewel, trust has many planes, each of which intersects with others. Only together can they create a brilliant image, and cracks in any single plane can threaten the whole. On the simplest level, people will have to be coaxed into using these new-fangled 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.[9] 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.[10]

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

[9] *See, e.g.*, NAT'L HIGHWAY TRANSP. SAFETY ADMIN., VEHICLE-TO-VEHICLE COMMUNICATION, (2018), www.nhtsa.gov/technology-innovation/vehicle-vehicle-communication.

[10] *See* Chetan Belagal Math et al., *Data Rate based Congestion Control in V2V communication for traffic safety applications*, 2015 IEEE Symposium on Communications and Vehicular Technology in the Benelux (2015), www.semanticscholar.org/paper/Data-Rate-based-Congestion-Control-in-V2V-for-Math-Ozgur/72570d8a3be7bbe501e68a41e260862400cfbc35; Mary Beth Griggs, *The Safer, Faster, More Efficient Commute Of The Future*, POPULAR SCIENCE (Feb. 26, 2015), www.popsci.com/safer-faster-more-efficient-commute-future.

uncertainty of what the human driver will choose to do.[11] In fact, in the current tests of driverless car systems, some of the greatest difficulties flow from interacting with human drivers on the road 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.[12]

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 – deep learning models being especially opaque. When decisions are being made that result in sending criminals to jail[13] or choosing between killing the driver of an autonomous vehicle and a crowd of six,[14] how do we develop the pathways for interrogating the technology to society's satisfaction? 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 circum-

[11] *See* Matt Richtel & Conor Dougherty, *Google's Driverless Cars Run into Problems: Cars with Drivers*, N.Y. TIMES (Sept. 1, 2015), www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html?_r=0.

[12] Kristen Hall-Geisler, *All new cars could have V2V tech by 2023,* TECHCRUNCH (Feb. 2, 2017), techcrunch.com/2017/02/02/all-new-cars-could-have-v2v-tech-by-2023/.

[13] *See* Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, *Machine Bias*, PROPUBLICA (May 23, 2016), www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[14] *See* Matt Simon, *To Make Us All Safer, Robocars Will Sometimes Have to Kill,* WIRED (March 13, 2017), www.wired.com/2017/03/make-us-safer-robocars-will-sometimes-kill/.

stance, the breakdown of trust can be more serious than the disarray that can develop when individuals opt out of a linked network.

Just as a failure of trust can be detrimental, we also cannot afford to indulge in a failure of skepticism about AI. As described earlier, the optimal state of AI systems for the foreseeable future will involve human enhancement systems, that is, systems that work hand-in-bolt with humans. These may process information more quickly or more thoroughly. They may also enhance perception and reaction; but the systems will need human interface and, most important, human redundancy to provide the type of analysis and confirmation that only humans can.

At a simplistic level, society will need to guard against the type of overconfidence that will lead us to attribute unrealistic capacity and accuracy to specific AI systems. Tesla drivers, for example, have climbed into the backseat of the car or driven with newspapers in front of their faces, giving the vehicles more unsupervised rein than their current capacity merits.[15] As AI systems reach into everything from legal decisions to labor to health care, our expectations must keep pace with their limitations.

Human judgment and interpretation also will be important for responding to and recovering from the types of security incursions that artificial technologies will face. Although most people think of network security in terms of warding off attacks and preventing penetration, the security field is moving towards a greater emphasis on detection and recovery. Estimates of the number of personal records stolen in 2016 alone are in the billions, and cyberattacks against the U.S. government have increased over the last decade from 5,500 to 77,000 a year.[16] 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 an attacker has gained access to the network or even after a successful attack is underway.

[15] *See* Mike Ramsey, *Driver Videos Push Tesla's 'Autopilot' to Its Limits,* WALL ST. J. (Oct. 25, 2015), www.wsj.com/articles/driver-videos-push-teslas-autopilot-to-its-limits-1445765404.

[16] *See* GOV'T ACCOUNTABILITY OFF., AGENCIES NEED TO IMPROVE CONTROLS OVER SELECTED HIGH-IMPACT SYSTEMS*,* (May 18, 2016), www.gao.gov/products/GAO-16-501; FIRE-EYE, M-TRENDS 2015: A VIEW FROM THE FRONTLINES, www2.fireeye.com/rs/fireye/images/rpt-m-trends-2015.pdf.

When systems are extensively networked, infection of a single weak point can have widespread consequences. In the summer of 2017, for example, the "NotPetya" malware attack operated by accessing a popular accounting firm and then inserting malware that spread throughout millions of computers when users updated their accounting software.

AI systems will be no exception. Consider automobiles. When networked, driverless cars become the norm, the vulnerable entry points for the system 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. As a hacker, I only need to 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. For example, driverless cars use machine learning to decide what objects are in their surroundings, including whether something is a stop sign or a speed limit sign. By poisoning the data, one can corrupt the car's decision making. One study, albeit a limited one, managed to confuse automated systems about the nature of stop signs by placing unobtrusive stickers on the signs.[17] In other words, one can wreak havoc either by poisoning the training data or by poisoning the input data, if the input data is sufficiently open.

With wearable technologies, particularly those related to health, the dangers can be particularly troubling. Research already has shown that heart pacemakers can be hacked.[18] The more wearable and implantable technologies spread throughout society, the greater the consequences of tampering.

The mortal consequences make either human supervision, or a degree of conservatism, a necessity. If an AI boss system is compromised and assigns

[17] *See* Jonathan M. Gitlin, *Hacking Street Signs with Stickers Could Confuse Self-Driving Cars,* ARSTECHNICA (Sept. 1, 2017); *see also* Tianyu Gu et al., *BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain,* THE MORNING PAPER (Oct. 13, 2017), blog.acolyer.org/2017/10/13/badnets-identifying-vulnerabilities-in-the-machine-learning-model-supply-chain/.

[18] Natt Garun, *Almost Half a Million Pacemakers Need a Firmware Update to Avoid Getting Hacked*, THE VERGE (Aug. 30, 2017), www.theverge.com/2017/8/30/16230048/fda-abbott-pacemakers-firmware-update-cybersecurity-hack.

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 cyber security project, can be detected by a human thinking "someone else could do a better job at this."

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 then puzzle through.

Perhaps the greatest potential risk with widespread adoption of AI lies with circumstances in which disruptions are combined with fear. In that context, small incursions can have an echo effect, magnifying the harm exponentially. 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.[19] 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. These examples are somewhat analogous to terrorism attacks. The death toll from large-scale terrorist attacks in this country in the last year was small compared with the death toll from "boring" killers, like the flu. In fact, from 12,000 to 56,000 people have died from the flu each year in the United States since 2010.[20] Nevertheless, terrorism causes panic, underlying anxiety, and an erosion of trust in a manner that is widespread and culturally significant.

In this way, trust and distrust can wrap back around each other and collide to provide the maximum risk for chaos and societal disruption. Imagine a time in which each person has an implanted health device, call it

[19] *See text accompanying notes* 22-23 *infra* (describing the fallout from a 1982 episode of malicious tampering with Tylenol bottles).

[20] CENTERS FOR DISEASE CONTROL AND PREVENTION, ESTIMATING SEASONAL INFLUENZA-ASSOCIATED DEATHS IN THE UNITED STATES, www.cdc.gov/flu/about/disease/us_flu-related_deaths.htm (last updated Jan. 29, 2018).

a health regulator. The device contains that person's health information, monitors various bodily functions, and can even direct implantable devices such as pacemakers or automated medicine dispensing mechanisms. The technology for such a device – at least a rudimentary one – is not too far off in the future. Now consider the following hypothetical. 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. A nurse recognizes the information is suspect, given that the blood type AB negative is almost nonexistent in the Korean population. Further investigation shows that the person's health data has been corrupted and that the problem most likely extends beyond this one patient. Early indications suggest that the health data problem exists only in relation to recent immigrants from North Korea, and that it was done intentionally.

The fallout from such a data incursion into a health care network could be extensive. Patients from North Korea might refuse to receive medical treatment, for fear that their health information would lead the medical team astray. Those fears could cascade throughout immigrant populations, or throughout patient populations in general, both for rational and irrational reasons. On the irrational side, the public in general or smaller groups in particular could easily misunderstand whether the data corruption extends beyond the North Korean immigrant population. On the rational side, citizens might fear that a limited incursion could be the beginning of larger incursions, either by this attacker or by others.

On another level, a 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. Should a medical care facility move to hand checking information for all patients, a laborious process, or only recent North Korean immigrants? How will North Korean immigrant populations respond if delivery of their health services is slowed in comparison to health services for others?

The potential social implications also are profound. Disruption of the health care 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 sources of information and not all of those sources would necessarily have the best of intentions.[21]

Looking beyond civilian implications, government actors would be pressed to determine whether the data corruption represents a ransomware or malware attack that merely uses immigrant populations as a convenient entry point; sabotage from North Korea; a domestic player targeting immigrants; or even enemies of North Korea trying to sow distrust. Regardless of the source, the attack could represent one prong of a larger campaign or simply provide inspiration for others. In short, a small and limited incursion could have extensive and profound effects on social cohesion and societal resources.

III.
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.[22] 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.

[21] *See, e.g.,* Jiayang Fan, *Chinatown's Ghost Scam: When Elderly Immigrants Fall Prey to Fraudsters Promising Protective Blessings, Their Life Savings Are Spirited Away,* NEW YORKER (Oct. 30, 2017), www.newyorker.com/magazine/2017/10/30/chinatowns-ghost-scam.

[22] *See* Ronald Reagan, Pres., U.S., Statement on Signing the Federal Anti-Tampering Act (Oct. 14, 1983) (describing the Tylenol attack's relationship to the legislation, and noting that "every American became keenly aware of the tremendous harm that can be done by a single deranged person"), www.presidency.ucsb.edu/ws/index.php?pid=40636.

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. (And of course, as the techie saying goes, weeks of programming can save hours of planning.) 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,[23] 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.

In particular, 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. For example, science creates pathways that drive legal regimes, because law cannot dictate what science cannot accomplish.[24] In turn, law affects the unfolding of scientific development, and not simply by holding out the promise of goodies such as intellectual property rights or punishments such as legal liability. Law also shapes the expectations of individual citizens, developing them and trimming them. When a car driven by a 16-year-old hits mine on the road, I expect the driver 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 trick, and it will indeed be tricky, will be to ensure that as these technologies

[23] *See* Robin Feldman, *Coming to the Community, in* IMAGINING NEW LEGALITIES, AMHERST SERIES IN LAW, JURISPRUDENCE, AND SOCIAL THOUGHT 88 (Austin Sarat ed., Stanford 2012) (describing Lockean legal theories); *see also* Peter Laslett, *John Locke: Two Treatises of Government: A Critical Edition* 336-37, 341, 343 (Cambridge 1967) (including a reprinting of the 1698 version of Locke's *Two Treatises*) (describing the individual's voluntary transition from "an uncontrolled enjoyment of all the Rights and Privileges of the Law of Nature" to membership in a political society in which "he authorizes the Society, or which is all one, the Legislative thereof to make Laws for him as the public good of the society").

[24] *Cf.* LAWRENCE LESSIG, CODE & OTHER LAWS OF CYBERSPACE (Basic Books 1999) (arguing that software and hardware regulate liberty in cyberspace in a manner analogous to legal regulations).

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 also fall short where openness is the goal. Although patents, in theory, disclose sufficient information so that others can make and use the invention, the reality is quite different. 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.[25] Some other forms of invention rights might be needed.

In addition, AI systems should be subject to review entirely outside the system itself – either 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 on a broad scale. Think about the quintessential notion of the distribution of power among the states within the federal system. The electoral college – much maligned in modern commentary – played a central role in ensuring the coalescence of a nation in which small and sparsely populated states feared domination by the mighty few.[26] This is not to suggest that the arrival of AI lies on a par with the birth of the nation. Rather, the juxtaposition of imagery is a reminder of the power of considered thought in contrast to frenzied reaction. We are at the dawn of an era, albeit one that is scientific rather than political. Rather than stumbling blindly, we should move with great care and intention.

[25] *See* ROBIN FELDMAN, RETHINKING PATENT LAW 104-127 (Harvard 2012) (describing the development of modern patent doctrines for software-related inventions).

[26] *See generally* William C. Kimberling, Deputy Dir., Fed. Elec. Comm'n Off. of Elec. Admin., *The Electoral College*, transition.fec.gov/pdf/eleccoll.pdf (revised May 1992).