2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data

JACK M. BALKIN*

TABLE OF CONTENTS

I. THE FRANKENSTEIN COMPLEX

When I was a boy, I read all of Isaac Asimov's stories about robotics. In Asimov's world, robots were gradually integrated into every aspect of society. They had various degrees of similarity to humans, but as the stories and novels progressed, the most advanced robots were very human in appearance and form.

The most famous feature of these robot stories is Asimov's three laws of robotics that were built into every robot's positronic brain.

The three laws are:

First Law: "a robot may not injure a human being, or, through inaction, allow a human being to come to harm."[1]

Second Law: "a robot must obey the orders given it by human beings except where such orders would conflict with the First Law."[2]

Third Law: "a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws."[3]

---

* Knight Professor of Constitutional Law and the First Amendment; Director, The Information Society Project, Yale Law School. This Lecture was originally presented as the 2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy at the Ohio State University Moritz College of Law on October 27, 2016. My thanks to Dennis Hirsch, Margot Kaminski and Frank Pasquale for their comments.

[1] ISAAC ASIMOV, *Runaround, in* I, ROBOT 41, 53 (Gnome Press 1st ed. 1950).

[2] *Id.*

[3] *Id.* at 54.

These three laws have been very influential, and even today people imagine what it would be like—or whether it would even be possible—to build them into robots, including, for example, into self-driving cars.[4]

As a dramatic device, the laws of robotics replaced one familiar trope about robots with a far more interesting one. The older trope was the idea of the Frankenstein monster or the killer robot, which becomes evil or goes berserk. An example of this literary theme in the Terminator movie franchise is the neural network Skynet becoming self-aware and taking over the world.[5] But Asimov wrote his robot stories to counteract what he called the "Frankenstein Complex"—the idea that robots were inherently menacing or evil, and that human beings would inevitably create mechanical beings who would turn on their creators.[6] In many of his stories, in fact, people start out as prejudiced against robots and then end up seeing their value. For example, the protagonist of several of his stories, Detective Elijah Bailey, who is initially skeptical of robots, eventually becomes best friends with R. Daneel Olivaw, his robotic partner.[7]

By creating the three laws, Asimov made things much more interesting. Instead of just worrying about whether robots would eventually turn on us, he raised an additional problem that is very near and dear to lawyers—this is the problem of legal interpretation. What do we—or in some cases, the robots themselves—do when the laws are unclear, or when they conflict? By creating the three laws, Asimov moved our imagination about robots from threats to objects of interpretation and regulation, and thus sources of irony and conflict. It is a very sophisticated idea, and one that he develops in many of his stories.

Today, it's quite unclear whether we could actually build the three laws that Asimov postulated into robots and artificial intelligence (AI) agents. After all, Asimov's three laws seem rather vague and incomplete. They might have loopholes.

---

[4] *See* Boer Deng, *Machine Ethics: The Robot's Dilemma*, NATURE (July 1, 2015), http://www.nature.com/news/machine-ethics-the-robot-s-dilemma-1.17881 [https://perma.cc/5HT6-EEFA] (relating Asimov's three laws to the developing field of machine ethics); *cf.* Ulrike Barthelmess & Ulrich Furbach, *Do We Need Asimov's Laws?*, CORNELL U. LIB. 11 (2014), https://arxiv.org/ftp/arxiv/papers/1405/1405.0961.pdf [https://perma.cc/YSC2-WZYJ] (arguing that the laws reflect cultural anxieties about robots, and it is unnecessary to build the laws into actual robots).

[5] THE TERMINATOR (Hemdale Film Corp. 1984).

[6] Isaac Asimov, *The Machine and the Robot*, *in* SCIENCE FICTION: CONTEMPORARY MYTHOLOGY 244, 250–53 (Patricia Warrick et al. eds., 1978); Lee McCauley, *The Frankenstein Complex and Asimov's Three Laws*, *in* HUMAN IMPLICATIONS OF HUMAN-ROBOT INTERACTION 9, 9–10 (Ass'n for the Advancement of Artificial Intelligence, Workshop Technical Report No. WS-07-07, 2007), https://www.aaai.org/Papers/Workshops/2007/WS-07-07/WS07-07-003.pdf [https://perma.cc/AJ9F-FHTJ].

[7] *E.g.*, ISAAC ASIMOV, THE CAVES OF STEEL 13–14, 190–91 (1954); ISAAC ASIMOV, *Mirror Image*, *in* ROBOT VISIONS 319, 319 (1990); ISAAC ASIMOV, THE NAKED SUN 19–20 (Harper Collins 1996) (1957); ISAAC ASIMOV, THE ROBOTS OF DAWN 30–31 (1983).

Of course, that was part of the point. A recurring trope in Asimov's stories is that the three laws are unclear, or vague, or might conflict in certain circumstances. Thus, the plot often turned on clever ways to interpret or reinterpret the laws of robotics, or on how to resolve conflicts between them, and so on. And in a late novel, Daneel Olivaw, a robot who begins as a detective but ends up being a very important figure in the novels, becomes so advanced that he creates his own "zeroth" law—"A robot may not injure humanity or, through inaction, allow humanity to come to harm"—that precedes all of the others that he received in his original programming.[8]

In any case, my goal is to ask how we might use Asimov's idea of the laws of robotics today. When I talk of robots, however, I will include not only robots—embodied material objects that interact with their environment—but also artificial intelligence agents and machine learning algorithms. That is perfectly consistent with Asimov's concerns, I think. Although Asimov wrote primarily about robots, he also wrote about very intelligent computers.[9] And the Frankenstein syndrome that he was trying to combat could arise from fear of AI or algorithms as much as fear of embodied robots. Today, people seem to fear not only robots, but also AI agents and algorithms, including machine learning systems.[10] Robots seem to be just a special case of a far larger set of concerns.

We are rapidly moving from the age of the Internet to the Algorithmic Society, and soon we will look back on the digital age as the precursor to the Algorithmic Society. What do I mean by the Algorithmic Society? I mean a society organized around social and economic decision-making by algorithms, robots, and AI agents, who not only make the decisions but also, in some cases, carry them out. The use of robots and AI, therefore, is just a special case of the Algorithmic Society.

Big Data, too, is a feature of the Algorithmic Society. In fact, Big Data is just the flip side of a society organized around algorithmic decision-making. Big Data is the fuel that runs the Algorithmic Society; it is also the product of its operations. Collection and processing of data produces ever more data, which

---

[8] ISAAC ASIMOV, ROBOTS AND EMPIRE 291 (1985); *see also* JOSEPH A. ANGELO, ROBOTICS: A REFERENCE GUIDE TO THE NEW TECHNOLOGY 103 (2007).

[9] *E.g.*, ISAAC ASIMOV, *The Last Question*, *in* ROBOT DREAMS 220, 220 (1986).

[10] *See, e.g.*, Rory Cellan-Jones, *Stephen Hawking Warns Artificial Intelligence Could End Mankind*, BBC (Dec. 2, 2014), http://www.bbc.com/news/technology-30290540 [https://perma.cc/JJE2-95RL]; Samuel Gibbs, *Elon Musk: Artificial Intelligence Is Our Biggest Existential Threat*, GUARDIAN (Oct. 27, 2014), https://www.theguardian.com/technology/2014/oct/27/elon-musk-artificial-intelligence-ai-biggest-existential-threat [https://perma.cc/KGV3-GAV5] ("With artificial intelligence we are summoning the demon."). Roboticists and AI researchers, on the other hand, may support specific forms of regulation but tend to be less frightened. *See, e.g.*, Connie Loizos, *This Famous Roboticist Doesn't Think Elon Musk Understands AI*, TECH CRUNCH (July 19, 2017), https://techcrunch.com/2017/07/19/this-famous-roboticist-doesnt-think-elon-musk-understands-ai/ [https://perma.cc/TC7R-NUK5] (quoting Rodney Brooks as pointing out that the famous people like Musk who are most concerned about artificial intelligence "don't work in AI themselves," while "regulation on self-driving Teslas . . . [is] a real issue").

in turn, can be used to improve the performance of algorithms.[11] To vary Kant's famous dictum, algorithms without data are empty; data without algorithms are blind.[12]

In this Lecture, I'm going to offer three new laws of robotics for the Algorithmic Society. In the process, I will also introduce four important theoretical ideas that will help us understand how we should regulate these entities. The four ideas are (1) the *homunculus fallacy*; (2) the *substitution effect*; (3) the concept of *information fiduciaries*; and (4) the idea of *algorithmic nuisance*. I'll explain these four ideas as the Lecture progresses.

Although I am inspired by Asimov's three laws of robotics, I nevertheless will describe the idea of "laws of robotics" very differently than he did.

First, these laws will not be limited to robots—they will apply to AI agents and algorithms, including machine learning algorithms. And when I want to talk about all three together as a group, I will talk about the laws of algorithms generally.

Second, when people think about robots in science fiction, they often think of self-contained entities. But today we know that many robots and AI agents are connected to the cloud.[13] That is certainly true of the Internet of things and home robots. It is likely to be true of self-driving cars as well. So the laws of robotics, whatever they are, are also likely to be the laws of cloud intelligences that are connected to the Internet.

Third, because robots are cloud robots, we shouldn't forget that one of the central issues in the study of robotics and artificial intelligence is the handling of data and, in particular, Big Data. Robots are nothing without data; and because many robots will be cloud robots, and many AI systems will be

---

[11] *See, e.g.*, *Fuel of the Future: Data Is Giving Rise to a New Economy*, ECONOMIST (May 6, 2017), https://www.economist.com/news/briefing/21721634-how-it-shaping-up-data-giving-rise-new-economy [https://perma.cc/H6RR-FZU2] ("Data will be the ultimate externality: we will generate them whatever we do."); Amir Gandomi & Murtaza Haider, *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L J. INFO. MGMT. 137, 140 (2015), http://www.sciencedirect.com/science/article/pii/S0268401214001066 [https://perma.cc/W3DL-YC5R] (describing how companies use algorithms to turn vast amounts of unstructured data into new forms of data that in turn can be used for analysis and decision-making); SINTEF, *Big Data, for Better or Worse: 90% of World's Data Generated over Last Two Years*, SCIENCEDAILY (May 22, 2013), https://www.sciencedaily.com/releases/2013/05/130522085217.htm [https://perma.cc/2G8V-ZYKE] ("Each and every one of us is constantly producing and releasing data about ourselves. We do this either by moving around passively—our behaviour being registered by cameras or card usage—or by logging onto our PCs and surfing the net.").

[12] IMMANUEL KANT, CRITIQUE OF PURE REASON 193–94 [A51/B76] (Paul Guyer & Allen W. Wood eds. & trans., Cambridge University Press 1998) ("Thoughts without content are empty, intuitions without concepts are blind.").

[13] *See* Ben Kehoe et al., *A Survey of Research on Cloud Robotics and Automation*, 12 IEEE TRANSACTIONS ON AUTOMATION SCI. & ENGINEERING 398, 400 (2015), http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7006734&tag=1 [https://perma.cc/DW7U-C4XU] ("The Cloud can provide robots and automation systems with access to vast resources of data that are not possible to maintain in onboard memory.").

connected to the Internet cloud, they will depend heavily on data analytics.[14] Data is the fuel that drives the engines of artificial intelligence.

So when we talk about robots, AI agents, and algorithms, we are also usually talking about Big Data and Internet connection, just as when we talk about Big Data, we are also usually talking about the regulation of robots, algorithms and AI agents that process it. The laws of robotics are also the laws of robotics, algorithms, and AI in the age of Big Data. Hence, the title of this Lecture.

Fourth, and perhaps most important, Asimov called his laws the laws of robotics, not the laws of robot-users or robot-programmers or robot-operators.[15] His laws were robot-directed.[16] They were robot-centric—that is, they were programming instructions inserted into the code of the robots themselves. They were laws that robots had to follow—because they were programmed that way—and not that users of robots had to follow.[17] You can imagine such instructions also being part of AI agents or algorithms as a check on the algorithm's machine learning. They are a sort of software side constraint, in the same way that rights are side constraints on action.

To be sure, humans were required to program the laws into every robot, but the laws themselves were addressed to the robots, not to the humans.[18] Asimov doesn't say much about the human laws that required this programming, but one assumes that there was some sort of government requirement that they be placed into every robot's positronic brain.

I will diverge from Asimov at this point. Instead of focusing on laws directed at robots (or algorithms), I focus on laws directed at the people who program and use robots, AI agents, and algorithms. That is because what we need in the emerging Algorithmic Society are not laws of robotics, but laws of robot-operators.

The conceit of the Algorithmic Society is the harnessing of data and algorithms to govern and improve society. The ambition of the Algorithmic Society is omniscience—to know all and to predict all—an ambition as old as humanity itself, but now seemingly ever closer to being within our grasp.

In the Algorithmic Society, the central problem of regulation is not the algorithms, but the human beings who use them, and who allow themselves to be governed by them. Algorithmic governance is the governance of humans by humans using a particular technology of analysis and decision-making.

---

[14] Bob Violino, *Big Data and Robotics: A Long History Together*, ZDNET (Aug. 12, 2016), http://www.zdnet.com/article/big-data-and-robotics-a-long-history-together/ [https://perma.cc/CA9W-C8VC] (noting that the concept of Big Data "has long been a part of the world of robotics," and that "[r]obotics was always about data"); Kehoe et al., *supra* note 13, at 401 (explaining how access to Big Data in the cloud enhances the ability of robots to perform tasks and interact with their environments).

[15] ASIMOV, *supra* note 1, at 53.

[16] *Id.*

[17] *Id.*

[18] *Id.* at 53–54.

Hence our need is not for robot-directed laws like Asimov's three laws of robotics, but laws directed at those who *use* robots to analyze, control, and exercise power over other human beings.

II. THE RABBI AND THE GOLEM

Let me explain this idea with a story. Margot Kaminski and I taught the first law and robotics course at Yale Law School in the spring of 2014. She came up with a reading list that I have used in subsequent years. For the first class, she chose a selection of famous literary examples, including a short story by Asimov which announces the three laws,[19] and Karel Čapek's 1921 play, R.U.R.,[20] which is the origin of the word "robot." She also chose a version of the legend of the Golem of Prague.[21]

According to legend, the Golem of Prague was created by Rabbi Judah Loew ben Bezalel, the Maharal, a sixteenth-century sage widely revered for his learning and piety.[22] The Maharal used the secret knowledge of Jewish mysticism to create a living thing out of clay, just as God had created Adam. He brought it to life by speaking the divine name. The Golem looked like a human. It was very strong, but it could not speak because, as the legend says, the power of speech was given to man alone by God[23] (the next time you talk to Siri, consider how things have changed).

In any case, the Maharal sets the Golem off to deal with threats to the Jewish community. In the legend, the Golem acts as a detective—just like Asimov's Daneel Olivaw. He finds out who is slandering the Jews, and he captures the bad guys. Then, having served his purpose, he returns to the Rabbi, who does the same secret incantations backwards, and the Golem turns back into a lifeless lump of clay, where he is stored in the attic of the synagogue.[24]

What is the point of this story? Well, the most interesting thing about this version of the story is what does not happen. The Golem doesn't go crazy. He doesn't catch the wrong person. The Rabbi's wife doesn't discover the Golem and accidentally set it loose; the Rabbi's son-in-law doesn't use the Golem to make money; an unscrupulous person doesn't retrain the Golem to do evil, and so on. In fact, nothing bad happens in this story. The Golem does exactly what it is supposed to do. And in a way, this version of the legend is rather boring;

---

[19] The story is *Runaround*. *Id.*

[20] KAREL ČAPEK, R.U.R. (1921).

[21] *The Golem of Prague*, *in* A TREASURY OF JEWISH FOLKLORE 603 (Nathan Ausubel ed., 1948).

[22] *Judah Loew ben Bezalel (The Maharal of Prague)*, HOLY PEOPLE OF THE WORLD: A CROSS-CULTURAL ENCYCLOPEDIA 450 (Phyllis G. Jestice ed., 2004) ("[T]he Maharal was respected as a learned and pious man by Jew and gentile alike."); JOAN COMAY, WHO'S WHO IN JEWISH HISTORY: AFTER THE PERIOD OF THE OLD TESTAMENT 208 (Oxford Univ. Press 2d ed. 1995) (noting that the Maharal was "[g]reatly revered for his piety and scholarship").

[23] *The Golem of Prague*, *supra* note 21, at 607–08.

[24] *Id.* at 609–11.

there are other versions of the Golem legend in which things go wrong, and it's far more interesting dramatically.[25]

But the most important lesson we can draw from this story is that the reason nothing goes wrong is that the Golem is programmed and employed by the Maharal, a man of the greatest piety and learning. Only a truly righteous man, or a saint, you might say, is capable of using the Golem only for good.

And this, in my view, is the real lesson of the story. When we talk about robots, or AI agents, or algorithms, we usually focus on whether they cause problems or threats. But in most cases, the problem isn't the robots; it's the humans.

Why is the problem the humans, and not the robots?

First, the humans design the algorithms, program them, connect them to databases, and set them loose.

Second, the humans decide how to use the algorithms, when to use them, and for what purpose.

Third, humans program the algorithms with data, whose selection, organization, and content contains the residue of earlier discriminations and injustices.

Fourth, although people talk about what robots did or what AI agents did, or what algorithms did, this way of speaking misses an important point. These technologies mediate social relations between human beings and other human beings. Technology is embedded into—and often disguises—social relations.

When algorithms discriminate or do bad things, therefore, we always need to ask how the algorithms are engaged in reproducing and giving effect to particular social relations between human beings. These are social relations that produce and reproduce justice and injustice, power and powerlessness, superior status and subordination.

The robots, AI agents, and algorithms are the devices through which these social relations are produced, and through which particular forms of power are processed and transformed.

This is what I mean when I say that the problem is not the robots; it is the humans.

III. THE HOMUNCULUS FALLACY

This brings me to the first of the four ideas that I promised I would talk about in this Lecture. I have coined a phrase—the *homunculus fallacy*—to describe the way that people tend to think about robots, AI agents, and algorithms. The homunculus fallacy is the belief that there is a little person inside the program who is making it work—who has good intentions or bad intentions, and who makes the program do good or bad things.

---

[25] *See, e.g.*, SHARON BARCAN ELSWIT, THE JEWISH STORY FINDER: A GUIDE TO 668 TALES LISTING SUBJECTS AND SOURCES 204–05 (2d ed. 2012) (listing the basic story and variations in Jewish literature).

But, in fact, there is no little person inside the algorithm. There is programming—code—and there is data. The program uses the data to run, with good or bad effects, some predictable, some unpredictable.

When we criticize algorithms, we are really criticizing the programming, or the data, or their interaction. But equally important, we are also criticizing the use to which they are being put by the humans who programmed the algorithms, collected the data, or employed the algorithms and the data to perform particular tasks. We are criticizing the Rabbi, not the Golem.

So what kinds of social relations do these technologies produce and reproduce? In order to explain this, I need to introduce the second of the four ideas in this Lecture, an idea which I've discussed previously. This is the *substitution effect*.[26]

The substitution effect refers to the effects on society that occur when robots, AI agents, and algorithms substitute for human beings, and operate as special purpose people.[27] The notion of robot or algorithm as substitute conveys four different ideas: (1) the substitute is in some ways better than the original; (2) the substitute is in other ways more limited than the original; (3) people treat the substitute as if it were alive—they engage in animism or anthropomorphism; and (4) the substitute acts as a fetish or deflection away from the social bases of power among human beings and groups of human beings.[28]

First, substitution means superiority: robots, AI agents, and algorithms are often more powerful and quicker than human beings and human decision-makers.[29] They can see things, do things, analyze things and make decisions that human beings could never do. They never tire of doing them, and they have no emotional distractions and no emotional compunction about doing them.[30]

Second, substitution also means limitation or deficiency. Robots, AI agents and algorithms have limited abilities. They can do only some things, but not others. They lack many of the features of human judgment.[31]

---

[26] Jack M. Balkin, *The Path of Robotics Law*, 6 CAL. L. REV. CIR. 45, 46, 55–59 (2015).

[27] *Id.*

[28] *Id.* at 57–59.

[29] *Id.* at 59.

[30] *Id.* at 58–59.

[31] Anupam Rastogi, *Artificial Intelligence—Human Augmentation Is What's Here and Now*, MEDIUM (Jan. 12, 2017), https://medium.com/reflections-by-ngp/artificial-intelligence-human-augmentation-is-whats-here-and-now-c5286978ace0 [https://perma.cc/3J7Q-FJJK] (noting that "[m]achines are in their relative infancy" in exercising common sense judgments that are easy for human beings, "in spite of rapid strides in Natural Language Processing using deep learning"); *see also* Catherine Havasi, *Who's Doing Common-Sense Reasoning and Why It Matters*, TECHCRUNCH (Aug. 9, 2014), https://techcrunch.com/2014/08/09/guide-to-common-sense-reasoning-whos-doing-it-and-why-it-matters/ [https://perma.cc/U938-MGJD] (noting that a central challenge of artificial intelligence research is developing capabilities for contextual, common-sense reasoning).

Third, substitution involves the projection of life, agency, and intention onto programs and machines.[32] This also encourages the projection of responsibility from the humans using the algorithms to the algorithms themselves—hence the homunculus fallacy.

Fourth, substitution involves a *fetish* or *ideological deflection*. Marx famously spoke of the fetishism of commodities.[33] Just as ancient societies believed that totems, which were inanimate objects, were imbued with magical powers, Marx argued that people in a market society treat the commodity as if it has value, when in fact what gives it value is the fact that it is embedded in a system of social relations.[34] Markets are social relationships that both empower people and allow people to exercise power over each other.

What is true of commodities in markets is also true of the use of technological substitutes in the form of robots, AI agents and algorithms. These technologies become part of social relations of power among individuals and groups. We must not confuse the Golem with the Rabbi. The effects of robotics are always about the relationships of power between human beings or groups of human beings.

Recently, the media reported a story about an algorithm for picking beauty contestants that preferred white people.[35] Such stories encourage the idea that algorithms have psychological biases. This is yet another example of the homunculus fallacy—there is no little beauty contestant judge inside the algorithm who is employing his or her prejudices. There is the history of previous beauty pageants, the cultural assumptions about beauty that inform these pageants, the kind of data that is collected, the way that the data is collected, the code that the algorithm employs, and the code for revising the code, if the algorithm employs machine learning. There are also the people who set the algorithm loose for a particular task. We must always remember that behind the Golem is the Rabbi (or a whole society of Rabbis) who make and use the Golem.

---

[32] Balkin, *supra* note 26, at 57.

[33] 1 KARL MARX, CAPITAL: A CRITIQUE OF POLITICAL ECONOMY 81 (Fredrick Engels ed., Samuel Moore & Edward Aveling trans., Charles H. Kerr & Co. 1909) ("The Fetishism of Commodities and the Secret Thereof.").

[34] *Id.* at 83 ("A commodity is therefore a mysterious thing, simply because in it the social character of men's labour appears to them as an objective character stamped upon the product of that labour.").

[35] Sam Levin, *A Beauty Contest Was Judged by AI and the Robots Didn't Like Dark Skin*, GUARDIAN (Sept. 8, 2016), https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people [https://perma.cc/46AK-D8CL]. A company called Beauty.AI devised the algorithm. *See Welcome to the First International Beauty Contest Judged by Artificial Intelligence: Beauty.AI 2.0*, BEAUTY.AI, http://beauty.ai/ [https://perma.cc/Z8H8-22WY].

IV. THE LAWS OF AN ALGORITHMIC SOCIETY

Let me summarize the argument so far. I began this Lecture with Asimov's three laws of robotics. I noted that these laws were laws directed to robots and to their code. Then, using the story of the Golem, I pointed out that the problem is not the robots, but the human beings. If so, then rather than Asimov's laws of robotics, what we really need are laws of robotics designers and operators. The laws of robotics that we need in our Algorithmic Society are laws that control and direct the human beings who create, design, and employ robots, AI agents, and algorithms. And because algorithms without data are empty, these are also the laws that control the collection, collation, use, distribution and sale of the data that make these algorithms work.

In sum, the laws of robotics that we need are laws governing the humans who make and use robots and the data that robots use.

What kinds of laws would these be? Return to my central point—that behind the robots, AI agents, and algorithms are social relations between human beings and groups of human beings. So the laws we need are obligations of fair dealing, nonmanipulation, and nondomination between those who make and use the algorithms and those who are governed by them.

People use algorithms to classify and govern populations of people. Because the relationship is one of governance, the obligations are fiduciary—of good faith, manipulation, and nondomination. These are the principles that should guide the Algorithmic Society. Unlike Asimov's three laws, these principles are not automatically built into the robots. We have to ensure that they characterize relationships between human beings. We must build them into our human society; we must program them into our laws.

What duties do algorithm users have toward society? To answer that question, consider the ambition of the Algorithmic Society. The dream of the Algorithmic Society is the omniscient governance of society.

From the ambitions come the harms. They include, in addition to the possibility of physical injury, violations of privacy, exposure, reputational harm, discrimination, regimentation (or normalization), and manipulation.

The Algorithmic Society is a way of governing populations. By governance, I mean the way that people who control algorithms analyze, control, direct, order, and shape the people who are the subjects of the data. People use algorithms to classify, select, comprehend, and make decisions about entire populations of people. This relationship is not simply a relationship of market profit. It is also a relationship of governance.

The Algorithmic Society also involves relationships of informational power. The AI knows a lot about you, but you don't know a lot about the AI. Moreover, you can't monitor very well what the AI agent or algorithm does. There is an asymmetry of power and an asymmetry of information between operators and those acted on or governed. This asymmetry is a central feature of the Algorithmic Society—it is the asymmetry of knowledge and of power

between the public and private governors of the Algorithmic Society and those who are governed by them.

What are the three laws—or more correctly, the legal principles—of the Algorithmic Society? They are three principles of fair governance.

(1) With respect to clients, customers, and end-users, algorithm users are *information fiduciaries*.

(2) With respect to those who are not clients, customers, and end-users, algorithm users have *public duties*. If they are governments, this follows from their nature as governments. If they are private actors, their businesses are affected with a public interest, as constitutional lawyers would have said during the 1930s.[36]

(3) The central public duty of algorithm users is to avoid externalizing the costs (harms) of their operations. The best analogy for the harms of algorithmic decision-making is not intentional discrimination, but socially unjustified pollution. Obligations of transparency, interpretability, due process and accountability flow from these three substantive requirements. Transparency—and its cousins, due process, accountability, and interpretability—apply in different ways with respect to all three principles.

Accountability, transparency, interpretability, and due process may be fiduciary obligations. They may follow from public duties. And they may be a prophylactic measure to prevent unjustified externalization of harms, or to provide a remedy for harm.

Let me discuss these three legal principles of robotics in turn.

V. FIRST LAW: ALGORITHMIC OPERATORS ARE INFORMATION FIDUCIARIES WITH RESPECT TO THEIR CLIENTS AND END-USERS

To discuss the first legal principle, I introduce yet another of the key ideas that I promised that I would mention in the course of this Lecture. This is the idea of *information fiduciaries*, a concept that I've developed in previous work.[37] To understand what an information fiduciary is, we should first ask, what is a fiduciary? Examples of fiduciaries are professionals like doctors and lawyers, and people who manage estates or other people's property.[38] What makes someone a fiduciary is that people depend on them to provide services, but there is a significant asymmetry in knowledge and ability between fiduciary and client. The client is in a position of special vulnerability, and can't easily monitor what the fiduciary is doing on his or her behalf.[39] As a result, the law

---

[36] *See* Nebbia v. New York, 291 U.S. 502, 536 (1934) ("The phrase 'affected with a public interest' can, in the nature of things, mean no more than that an industry, for adequate reason, is subject to control for the public good.").

[37] *See generally* Jack M. Balkin, *Information Fiduciaries and the First Amendment*, 49 U.C. DAVIS L. REV. 1183 (2016) (introducing and explaining the concept of an information fiduciary).

[38] *Id.* at 1207.

[39] *Id.* at 1216–17.

requires fiduciaries to act in a trustworthy manner, in good faith, and to avoid creating conflicts of interest with the client or patient.[40] Fiduciaries often collect sensitive personal information about their clients, which they could use to their client's detriment. Hence the law requires them to protect their client's privacy and not to disclose information in ways that would harm their clients.[41] When fiduciaries collect and process information about their clients, we can give them a special name. They are information fiduciaries.[42] Most professionals who are fiduciaries are also information fiduciaries.[43]

Fiduciaries have two central duties. The first is a duty of care.[44] The second is a duty of loyalty.[45] The duty of care means that a fiduciary has to act with reasonable care to avoid harming the client or patient.[46] The duty of loyalty means that the fiduciary has to avoid creating conflicts of interest with their clients or patients and must look out for their interests.[47] The degree of loyalty demanded depends on the nature of the relationship between the fiduciary and the client.

The digital age has created a new set of entities that have many features similar to traditional fiduciaries. They include large online businesses like Google, Facebook, and Uber. These businesses collect, collate, analyze, and use information about us.[48] Indeed, they collect enormous amounts of information about us, which could, in theory, be used to our detriment. These businesses have become quite important, in some cases indispensable, to our everyday lives. There is also an asymmetry of knowledge between businesses and their end-users and clients. Online businesses know a lot about us;[49] we know comparatively little about their operations, and they treat their internal processes as proprietary to avoid theft by competitors.[50] At the same time, these businesses attempt to reassure their end-users that they will respect their privacy and will not betray their trust.[51] Because they are, in Frank Pasquale's terms, a black box,[52] most people simply have to trust them.

---

[40] *Id.* at 1207–08.

[41] *Id.*

[42] *Id.* at 1208.

[43] Balkin, *supra* note 37, at 1209.

[44] *Id.* at 1207–08.

[45] *Id.* at 1208.

[46] *Id.* at 1207–08.

[47] *Id.* at 1208.

[48] *See generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 58–100 (2015) (discussing the secrecy involved in corporate collation, analysis, and use of personal data).

[49] *Id.*

[50] *Id.*

[51] *See, e.g.*, *Privacy Policy*, SNAP INC. (June 5, 2017), https://www.snap.com/en-US/privacy/privacy-policy/ [https://perma.cc/P99H-75JT].

[52] *See generally* PASQUALE, *supra* note 48.

I argue that businesses like these have many of the trappings of traditional fiduciaries.[53] They collect information about us. They watch us, but we cannot easily watch them; we become dependent on them and vulnerable to them, and so we have to trust them.[54] Such relationships traditionally have led to fiduciary status.[55] Hence I argue that these businesses should have legal obligations to be trustworthy toward their end-users. They are the digital age versions of information fiduciaries.[56]

Nevertheless, the duties of digital age information fiduciaries are different from those of doctors and lawyers. They are more limited because of the kind of services they render and because of the kinds of reasonable trust they create.[57]

First, unlike doctors and lawyers, monetizing personal data is central to many online service companies, because it allows them to subsidize the services they provide or to offer them for free. Merely recouping expenses or making a profit from such information doesn't by itself violate their fiduciary duty.[58]

Second, many online providers, like search engines and social media sites, make money because end-users produce a constant stream of content and links. Thus, unlike traditional professionals, these companies have an interest in getting people to reveal as much about themselves as possible—or otherwise express themselves in public as much as possible—so that their activities will generate content and data that companies can index and analyze.[59]

Third, people expect doctors to do far more than merely not harm them; people also expect that doctors will look out for their interests and warn them about potential risks to their health, their diet, and so on. People do not expect such comprehensive obligations of care from their ISP's, search engines, and social media sites.[60]

Because of these differences, digital information fiduciaries should have different and fewer obligations than traditional professional fiduciaries like doctors, lawyers, and accountants. They are special-purpose information fiduciaries, and the kinds of duties that it is reasonable to impose on them should depend on the nature of the services they provide.

The central obligation of digital information fiduciaries is that they cannot act like con artists—inducing trust in their end-users to obtain personal information and then using that information in ways that betray that trust and work against the interests of their end-users.[61] Online businesses should not be able to hold themselves out as providing digital safety and respecting digital privacy and then manipulate and discriminate against their end-users; nor should

---

[53] Balkin, *supra* note 37, at 1221–22, 1228.
[54] *Id.* at 1207, 1222.
[55] *Id.* at 1207.
[56] *Id.* at 1221.
[57] *Id.* at 1225–26.
[58] *Id.* at 1225–27.
[59] Balkin, *supra* note 37, at 1225–27.
[60] *Id.* at 1226–27.
[61] *Id.* at 1224–25.

they be able to sell or distribute data about their end-users to companies who will not abide by similar duties of care and good faith.[62]

Currently, law does not treat these digital businesses like fiduciaries. But I have argued that it should. The law should extend fiduciary obligations to these companies and clarify the duties that online firms owe to their customers and end-users.[63]

Now think about home robots and smart houses. Home robots and smart houses collect an enormous amount of information about us which, in theory, can be collated with information about many other people that is stored in the cloud. Home robots and smart houses, in other words, aren't simply stand-alone products. They are always-on, interconnected cloud entities that rely on and contribute to huge databases. Although we may come to trust the home robot and the smart house—indeed, we have to—the entity that we really have to trust is not the robot or the house. It is the company behind the robot and the house that collects the data from the robot and from the house's sensors. And that company, I argue, should be an information fiduciary.

The owner of the fiduciary duty, in other words, is not the robot. It is the company that manufactures, installs, sells and operates the robot in our home. It is the Rabbi, and not the Golem, that owes us fiduciary obligations.

The first law of robotics for the algorithmic age, therefore, is that those who develop and employ robots, AI agents, and algorithms have duties of good faith and trust toward their end-users and clients. Fiduciary duties apply whether a business or entity uses robots, AI agents, or machine learning algorithms in delivering services.

Home robots and smart homes are obvious examples. Other examples might be services like Airbnb, Uber, OKCupid, Match.com, and 23 and Me. What matters in each case is that the businesses induce trust and collect personal information about us and might use it in ways that betray our trust and/or create a conflict of interest.

Earlier I noted that the classic examples of fiduciary obligations arise in the professions. In fact, robots, AI agents, and algorithms are likely to be increasingly employed in the operations of traditional fiduciaries—doctors, lawyers, accountants, and money managers. The federal government recently issued new rules through the Labor Department that will treat investment

---

[62] *Id.* at 1224–25, 1227; *id.* at 1233 (arguing that digital information fiduciaries "may also have duties to ensure that, when they sell or convey this information to others, duties of non-disclosure and non-manipulation travel with the data").

[63] *Id.* at 1223–24, 1226–29. *See also* Jack M. Balkin & Jonathan Zittrain, *A Grand Bargain To Make Tech Companies Trustworthy*, ATLANTIC (Oct. 3, 2016), https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/ [https://perma.cc/R6G4-UVKC] (arguing for a new "Digital Millennium Privacy Act").

advisors who handle retirement accounts as fiduciaries.[64] These advisors, in turn, are increasingly turning to AI and algorithms to do their jobs.[65]

The idea of fiduciary obligations also extends to governments that use robots, AI agents, and algorithms in their everyday functions, including the delivery of social services. Governments have duties of care and loyalty toward the populations that they govern.

VI. SECOND LAW: ALGORITHMIC OPERATORS HAVE DUTIES TOWARD THE GENERAL PUBLIC.

Because governments have fiduciary obligations to the people they govern, governments and public entities who use algorithms are information fiduciaries toward the populations they govern.

What about private actors? Some private actors, as we have seen, are information fiduciaries toward their clients, patients, and end-users. But not every private online business that uses robots, AI agents, or algorithms is an information fiduciary. Perhaps equally important, fiduciary duties generally extend only to a business's clients and end-users, and not to the general public as a whole.[66]

So the idea of information fiduciaries is not sufficient to explain all of the various obligations of private companies that use algorithms, AI agents, and robots. Businesses that employ algorithms in their operations may still cause harms to people who are not their clients or customers, and with whom they have no contractual relationship.[67] Examples are employers who are deciding whether to hire people or loan money to them—that is, enter into contractual relations with them—and credit reporting companies, who create our online reputations that others will employ.[68]

If we simply excluded all of the businesses that affect people but have no contractual relationships with them, we would be replicating a problem that emerged at the beginning of the twentieth century. In a modern industrial economy, businesses generated mass-produced goods that were no longer sold to consumers who directly contracted with them.[69] Instead, a chain of

---

[64] Definition of the Term "Fiduciary"; Conflict of Interest Rule—Retirement Investment Advice, 81 Fed. Reg. 20,946, 20,997 (Apr. 8, 2016) (to be codified at 29 C.F.R. § 2510.3–21 (2016)).

[65] *See* Brian O'Connell, *Will Robo-Advisors Benefit from the Fiduciary Rule?*, THESTREET (Feb. 23, 2017), https://www.thestreet.com/story/14009692/1/will-robo-advisors-benefit-from-the-fiduciary-rule.html [https://perma.cc/VB8M-YUNA] ("[R]obo firms are in good position to step in and snap up clients left behind by larger firms, who may shift their business focus to larger, more affluent clients as they adjust to the new fiduciary rule.").

[66] Balkin, *supra* note 37, at 1232–34.

[67] *See infra* Part VII.

[68] Balkin, *supra* note 37, at 1232–33.

[69] *See, e.g.*, Kyle Graham, *Strict Products Liability at 50: Four Histories*, 98 MARQ. L. REV. 555, 566–67 (2014) (recounting the standard history).

intermediaries brought these goods to market.[70] Consumer protections based on privity of contract were unsuited to new economic realities. As a result, courts, beginning with Cardozo's famous 1916 decision in *MacPherson v. Buick Motor Co.*,[71] abolished the privity rule and held that manufacturers had public duties, not only to direct consumers who purchased the products from intermediaries, but also duties to their family members and to bystanders who were injured by defective products.[72]

If we are to articulate the rules of the Algorithmic Society, we need something like *MacPherson v. Buick Motor Co.* for the Algorithmic Society. That is, we need to recognize that the use of algorithms can harm not only the end-user of a service, but many other people in society as well. For example, Jonathan Zittrain has pointed out how Facebook might use its data on end-users to manipulate them in order to swing a national election.[73] If that were to happen, it would affect not only the people with Facebook accounts, but everyone in the country. Similarly, when companies use algorithms in high speed trading, they can precipitate a market crash that affects not only the people they trade with, but everyone in the country, and indeed, the world.

It follows then that companies owe duties to the public when they employ robots, AI agents, and algorithms. But we can't describe the duties that they owe the public in terms of breach of trust toward clients, patients, and end-users. If these duties are not premised on betrayal of trust, what are they based on? This brings me to the third law of the Algorithmic Society.

VII. THIRD LAW: ALGORITHMIC OPERATORS HAVE A PUBLIC DUTY NOT TO ENGAGE IN ALGORITHMIC NUISANCE

What do I mean by algorithmic nuisance? Here I make an analogy to private and public nuisances—smells, smoke, sounds, poisons, and especially pollution. Traditionally, these were harms associated with the use (and misuse) of real property, but the idea has expanded in recent times to include a wide range of harms.[74] A private nuisance imposes harm on the recognized legal interests of a

---

[70] *Id.*; *see also* Randy Knitwear, Inc. v. Am. Cyanamid Co., 181 N.E.2d 399, 402 (N.Y. 1962) ("The world of merchandising is . . . no longer a world of direct contract.").

[71] MacPherson v. Buick Motor Co., 111 N.E. 1050 (N.Y. 1916).

[72] *Id.* at 1053.

[73] Jonathan Zittrain, *Engineering an Election*, 127 HARV. L. REV. F. 335, 335–36 (2014), http://harvardlawreview.org/2014/06/engineering-an-election/ [https://perma.cc/F4WX-29B3]; Jonathan Zittrain, *Facebook Could Decide an Election Without Anyone Ever Finding Out*, NEW REPUBLIC (June 1, 2014), http://www.newrepublic.com/article/117878/information-fiduciary-solution-facebook-digital-gerrymandering [https://perma.cc/3FUN-W2K9].

[74] *See* RESTATEMENT (SECOND) OF TORTS § 821A cmt. b (AM. LAW INST. 1979) (describing nuisance as "human activity or a physical condition that is harmful or annoying to others").

relatively small group of people;[75] a public nuisance diffuses harm over an indefinite population, and it is up to the state authorities to decide whether to bring an action to abate the nuisance.[76] In the alternative, the government must decide whether to create a scheme of government regulation akin to consumer or environmental protection.[77]

Obviously, I am not claiming that algorithmic harms are nuisances in the traditional common-law sense of that term. In particular, I am not saying that algorithmic harms are nontrespassory invasions of the private use and enjoyment of real property.[78] Rather, I argue that the best way to think about these harms is by analogy to torts like nuisance.

Why do I analogize the harms caused by algorithms to nuisance? I do so for three reasons. The first is the homunculus fallacy. We can't argue that the algorithm itself has bad intentions. Rather, the algorithm is used by human beings who want to achieve some particular set of managerial goals, but in the process, end up harming various groups of people. Some of these victims are easy to identify, but the harms to others are more diffuse.

In essence, we are talking about the socially unjustified use of computational capacities that externalizes costs onto innocent others. In tort law,

---

[75] *Id.* § 821E & cmt. a ("The liability for private nuisance exists only for the protection of persons having 'property rights and privileges,' that is, legally protected interests, in respect to the particular use or enjoyment that has been affected.").

[76] *Id.* § 821B(1) ("A public nuisance is an unreasonable interference with a right common to the general public."); *id.* § 821C(1) (stating that public officials must bring suits to abate a public nuisance unless a private individual suffers a harm different in kind from that suffered by the general public). In the past twenty years, state attorneys general have attempted to expand the concept of public nuisance to create a remedy for mass torts, public health problems, and environmental pollution. *See, e.g.*, Connecticut v. Am. Elec. Power Co., 582 F.3d 309, 314–15 (2d Cir. 2009) (holding that states could sue power companies for excessive carbon dioxide emissions under federal common law of public nuisance); Richard O. Faulk & John S. Gray, *Alchemy in the Courtroom? The Transmutation of Public Nuisance Litigation*, 2007 MICH. ST. L. REV. 941, 943–44 (2007) ("American jurisprudence has been experiencing another 'assault upon the citadel' in suits against asbestos, gun, and former lead paint manufacturers . . . [using] the lesser-known tort of 'public nuisance.'"); Donald G. Gifford, *Public Nuisance as a Mass Products Liability Tort*, 71 U. CIN. L. REV. 741, 743–44 (2003) (describing recent use of public nuisance suits by both state officials and private litigants to recover damages allegedly arising from tobacco-related diseases, firearm violence, and childhood lead poisoning); *see also* Victor E. Schwartz et al., *Game Over? Why Recent State Supreme Court Decisions Should End the Attempted Expansion of Public Nuisance Law*, 62 OKLA. L. REV. 629, 629–31 (2010) (describing and criticizing the expansion of public nuisance law beyond its traditional doctrinal boundaries).

[77] *See* RESTATEMENT (SECOND) OF TORTS § 821B(2)(b) & cmt. c (AM. LAW INST. 1979) (noting that legislatures and administrative agencies may determine that certain conduct constitutes a public nuisance, thereby obviating the need for an additional showing of unreasonable interference).

[78] *Id.* § 821D & cmt. a (noting that private nuisance has traditionally been concerned with nontrespassory invasions of interests in the use and enjoyment of land). Public nuisance, by contrast, might be concerned with broader matters like public health, safety, or morals. *Id.* § 821B & cmt. b.

we might call this externalization of a nuisance, whether public or private.[79] And indeed, in a recent article on how to regulate algorithmic policing, Andrew Selbst has argued that the appropriate remedy is to require police departments to create discrimination impact statements akin to environmental impact statements.[80] What is characteristic about algorithmic discrimination, Selbst argues, is that it cannot easily be identified with malice or bad intentions, either on the part of the officers using the programs or the programs themselves.[81] The algorithm doesn't have intentions, wants, or desires. That is the homunculus fallacy. There is no little person inside the algorithm who is directing it. Hence it is useless to model the duty or liability of algorithm operators on a respondeat superior theory—you can't impute intentions, negligence, or malice from the algorithm to the operator, even—and especially—a self-learning algorithm.

Instead, we have to focus on the social effects of the use of a particular algorithm, and whether the effects are reasonable and justified from the standpoint of society as a whole. Instead of drawing analogies to criminal law or the law of disparate treatment in antidiscrimination law, the best analogies are to nuisance and environmental law.[82]

The second reason to analogize the problem to nuisance is that the harms of algorithms are matters of degree. In addition, the harms of algorithmic nuisance

---

[79] Critically surveying the new enforcement actions based on public nuisance, Keith Hylton argues that there is a coherent rationale for the expansion of public nuisance, albeit one not always reflected in the case law: "Nuisance law induces actors to choose socially optimal activity levels by imposing liability when externalized costs are far in excess of externalized benefits or far in excess of background external costs." Keith N. Hylton, *The Economics of Public Nuisance Law and the New Enforcement Actions*, 18 SUP. CT. ECON. REV. 43, 44 (2010).

[80] Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 49 GA. L. REV. (forthcoming 2017) (manuscript at 44), https://ssrn.com/abstract=2819182 [https://perma.cc/67NC-Y6VE].

[81] *Id.* at 32 ("The problems of discrimination in data mining, however, are not those of motive, conscious or unconscious."). Often the discriminatory effects arise from the way a particular problem is framed for the algorithm to solve, rather than from unconscious motivations on the part of the police or programmers. *Id.*

[82] *See* A. Michael Froomkin, *Regulating Mass Surveillance as Privacy Pollution: Learning from Environmental Impact Statements*, 2015 U. ILL. L. REV. 1713, 1715 (2015) ("Market failures, collective action problems, and especially information asymmetries—including, we have recently learned, a stunning lack of government transparency about domestic surveillance—characterize the current privacy crisis, much as they did the environmental problem in the 1960s."); Dennis D. Hirsch, *Protecting the Inner Environment: What Privacy Regulation Can Learn from Environmental Law*, 41 GA. L. REV. 1, 23 (2006) (arguing that "[t]he privacy injuries of the Information Age are structurally similar to the environmental damage of the smokestack era" because of negative externalities and problems of collective action); *cf.* Dennis D. Hirsch, *The Law and Policy of Online Privacy: Regulation, Self-Regulation, or Co-Regulation?*, 34 SEATTLE U. L. REV. 439, 465–66 (2011) (noting how co-regulatory models in environmental law may be relevant to privacy protection).

result from the cumulative effects of data collation, analysis, and decision-making on people's digital identities.

Selbst points out that the harms caused by algorithmic discrimination don't fit well into a simple binary categorization of yes or no—i.e., either you have discriminated or you have not.[83] Rather, there are inevitable tradeoffs in design and in how programmers formulate the problem an algorithm is being asked to solve.[84] It may be difficult to determine a baseline of nondiscriminatory action against which to measure the algorithm's operation, and it may be difficult (if not impossible) to isolate the effects of the algorithm's operations to a single cause.[85] Ultimately the relevant question is whether you have imposed too many unjustifiable costs on innocent third parties. Algorithmic discrimination, like pollution, is a matter of degree.

The third reason for the analogy to nuisance is that it helps us understand how the harms of the Algorithmic Society arise from cumulative decision-making and judgment by a wide range of public and private actors. Companies and governments use Big Data and algorithms to make judgments that construct people's identities, traits, and associations.[86] These digital constructions of identity and traits affect people's opportunities—to employment, credit, financial offers, and positions. They also shape people's vulnerabilities—to increased surveillance, discrimination, manipulation, and exclusion. Companies and governments collect data about people from multiple sources and process the data to make new data. In processing data and making decisions, companies and governments contribute to the cumulative construction of people's digital identities, traits, and associations, which, in turn, constructs people's future opportunities and shapes their vulnerabilities.[87]

Other companies build on these collections of data, scores, and risk assessments and on the resulting digital constructions of traits, associations, and identity.[88] Companies and governments employ all of this information creatively in ever new contexts of judgment, yielding ever new insights, judgments, and predictions. In this way, people's lives are subject to a cascade of algorithmic judgments that fashion identity, opportunities, and vulnerabilities over time. Imagine, if you will, your digital identity as an informational stream into which a collection of new judgments, scores, and risk assessments are constantly being tossed.

---

[83] Selbst, *supra* note 80, at 48–49.

[84] *Id.* at 47–48.

[85] *Id.* at 47–48; Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671, 677–92, 718–19 (2016).

[86] *See* ROBINSON + YU, KNOWING THE SCORE: NEW DATA, UNDERWRITING, AND MARKETING IN THE CONSUMER CREDIT MARKETPLACE 4–5 (Oct. 2014), https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf [https://perma.cc/2EZJ-BLT8].

[87] *Id.*

[88] *Id.*

As more and more businesses participate in the collective process of digital identity shaping, this cascade of judgments increasingly shapes people's lives. It may shift a wide range of socially unjustified costs onto people in the form of constricted opportunities and enhanced vulnerabilities. The concept of algorithmic nuisance tries to capture these effects on individuals as public and private actors toss more and more judgments into the stream of information that represents individuals and is used to judge, classify, and control them.

The central problem we face today, therefore, is not intentional discrimination, but cumulative harm to identity and opportunities. Using algorithms repeatedly and pervasively in areas like policing, employment, housing, and access to credit will have cumulative effects on populations, as decision-makers draw on multiple sources of data to construct people's digital identities and reputations.[89]

In some cases, the harm may be traceable to careless programming and operation—mistakes in code, unreasonable assumptions, or biased data. Or it may result from the unreasonable use of algorithms, data sources, previous classifications, and categorizations employed for new purposes for which they were not designed. But in many cases the programmer and user may be able to make a plausible claim that their initial model is reasonable, given the task at hand, the data analyzed, and the background assumptions of the model. Even so, handing off decision-making to algorithms over time will predictably cast a wide range of harms onto individuals and members of particular groups.

A central concern is how identity—the association of persons with positive and negative associations and traits—is constructed and distributed in the Algorithmic Society. Decision-makers economize on decision-making not only by making their own algorithmic judgments, but also by importing algorithmic judgments that other parties have made about people's attributes, trustworthiness, and reputation.[90] Credit scores are an obvious example, but they are only a primitive illustration of what an Algorithmic Society can accomplish over time.

Instead of starting from scratch by developing their own scoring algorithms, decision-makers can economize by using scores and judgments already created by other algorithms used in different contexts and for different purposes, and modifying and updating them to suit their needs.[91] Some of the most important insights of the Algorithmic Society come from reimagining how data collected for one purpose might be used to shed light on what seemed at first to be an unrelated phenomenon or problem.

---

[89] *Id.*

[90] *Id.* at 4, 5 fig. 1.

[91] *Id.* at 6; Lois Beckett, *Everything We Know About What Data Brokers Know About You*, PROPUBLICA (June 13, 2014), https://www.propublica.org/article/everything-we-know-about-what-data-brokers-know-about-you [https://perma.cc/3V72-XRWZ] (explaining how data brokers collect and sell information gathered from many different sources for many different purposes).

Firms specialize in collecting, collating, and distributing people's identities to other decision-makers, who add their decisions to a growing digital stream or dossier.[92] This means that people's identities—including the positive and negative characteristics attributed to them—are constructed and distributed through the interaction of many different databases, programs and decision-making algorithms. And in this way, people's algorithmically constructed identities and reputations may spread widely and pervasively through society, increasing the power of algorithmic decision-making over their lives. As data becomes a common resource for decision-making, it constructs digital reputation, practical opportunity, and digital vulnerability.

In this world, focusing on intentional tort or even on the negligent construction and supervision of algorithms may be inadequate. Instead, the best analogy in tort law theory may be to the social costs that arise from socially unjustified levels of activity. Increased activity levels produce increased social costs, even when an activity is conducted with due care. Even assuming that the firm exercises due care—which, of course, it may not—the cumulative effects of increased activity may nevertheless throw too much harm onto the rest of society.[93] These are characteristic situations of nuisance.

Increased activity levels and increased social costs may arise when firms adopt new technologies that allow them to increase their levels of activity.[94] In this case, the switch to a new technology—algorithmic decision-making—allows governments and businesses to make more decisions affecting more lives more pervasively and more cheaply. The Algorithmic Society increases the rapidity, scope, and pervasiveness of categorization, classification, and decision; in doing so it also increases the side effects of categorization, classification, and decision on human lives. These side effects are analogous to the increased levels of pollution caused by increased factory activity.

To be sure, it takes two to create an injury. The social costs of algorithmic decision-making also arise from the actions of injured parties. Perhaps, then, we should also give citizens incentives to expose themselves less to the side effects of algorithmic judgment. But, in the Algorithmic Society, injured parties cannot easily absent themselves from seeking jobs, housing, medical care, and participating in the quotidian features of everyday life. In the Algorithmic Society, people throw off the data that later will be employed to judge them simply by living in a digital world. Nor can people easily contract to avoid the harms of algorithmic judgment. The collective action problems are enormous, not to mention the costs of obtaining information about their situation. In the Algorithmic Society, people's digital identities are produced by many different actors, their digital identities flow to a wide range of decision-makers, and the decisions are made by entities about which people know little.

---

[92] Beckett, *supra* note 91.

[93] *See* Hylton, *supra* note 79, at 48.

[94] *Id.*

If we follow the analogy to nuisance then, the Third Law of Robotics is that algorithm operators have a duty to the public not to "pollute,"—that is, unjustifiably externalize the costs of algorithmic decision-making onto others. Just as the transformation to an industrial society predictably increased the amount of social pollution, the shift to an Algorithmic Society will predictably increase the side effects of data collection, computation, and algorithmic judgment.

What are these costs or harms? Consider some of the most common harms created by the Algorithmic Society. These harms, I should emphasize, are in addition to traditional physical harms, such as those created by self-driving cars or industrial robots, and the dignitary harms caused by surveillance and exposure:

(1) *Harms to Reputation*. There are two central ways that algorithms affect reputation: the first is *classification*; the second is *risk assessment*. Algorithms affect reputation by branding you and others like you as *risky*—in other words, what it means for you to be you is a certain kind of risk or propensity. The way that risk manifests will be different in different contexts. It might include the idea that you (or the people who live in a particular area) create a financial risk, an employment risk, a risk of committing a future crime, a risk of expending lots of social services, a risk of returning items or being a costly customer, a risk of wasting ad dollars because you won't buy anything, and so on. In this case, algorithmic harm is the imputation that you are a risky person, which is a kind of stigma.

Risk assessment usually accompanies classification; the algorithm affects your reputation by placing you in a category or class, which is not necessarily an assessment of risk. The algorithm constructs groups in which you are placed and through which you are known and therefore potentially acted upon. Classification can affect your reputation without an assessment of risk because it says what kind of person you are and who you are treated as equivalent to (and, implicitly, better than or worse than according to some metric).

(2) *Discrimination*. Because of the assessment of risk and/or because of the work of classification, the enterprise that employs the algorithm denies you opportunities that it offers others (a credit card, a loan, a job opportunity, a promotion); or it imposes special costs (susceptibility to stop and frisk, surveillance, higher prices, exclusion from gun ownership or access to air travel, etc.) that it does not impose on other people.

(3) *Normalization or Regimentation*. The algorithm causes you to internalize its classifications and assessments of risk, causing you to alter your behavior in order to avoid surveillance or avoid being categorized as risky. It causes you to alter your identity, behavior, or other aspects of personal self-presentation in order to appear less risky to the algorithm, or to fall into a

different category; in the alternative, you engage in behavior that the algorithm does not pay attention to.[95]

(4) *Manipulation.* Human beings and organizations can use algorithms to lead you and others like you to make (more or less) predictable choices that benefit the algorithm operator but do not enhance your welfare and may actually reduce your welfare. In addition, algorithmic analysis makes it easier for companies to discover which people are most susceptible to manipulation, and how they can most easily and effectively be manipulated.

(5) *Lack of Due Process/Transparency/Interpretability.* The algorithm makes decisions that affect your welfare in one of the ways noted above without transparency, interpretability, an explanation of inputs and outputs in layman's terms, an ability to monitor the algorithm's operations, a means of providing rebuttal, or a method of holding the algorithm and its operators accountable.[96]

We might sum up this discussion by saying that algorithms (a) construct identity and reputation through (b) classification and risk assessment, creating the opportunity for (c) discrimination, normalization, and manipulation, without (d) adequate transparency, accountability, monitoring, or due process.

How are these harms related to the idea of algorithmic nuisance? These harms are the side effects of computerized decision-making. They are the social costs of algorithmic activity.

The pervasive adoption of algorithms greatly increases the level and pervasiveness of computational decision-making in our lives. Algorithmic use saves decision-makers money, and thus may be reasonable from the standpoint of an individual firm or decision-maker, but in the process it may impose cumulative harms on individuals and groups.[97]

Imagine, for example, a set of algorithms used to identify prospective employees. The algorithm may be good enough to fill up the small number of available slots with qualified people, but it excludes a large number of people who would also be qualified. (In this case, we would say that it creates very few

---

[95] Note that by conforming their behavior to algorithmic judgment or hiding themselves from data collection and algorithmic analysis, potentially injured parties can alter their own activity levels, and thus reduce the total social costs of algorithmic judgment. As noted before, it takes two parties to create a harm. But this way of reducing costs begs an important question: one must first demonstrate that society may reasonably demand a particular form of social regimentation or social isolation from its citizens.

[96] There is now an important literature grappling with the problems of due process in algorithmic decision-making and how best to address them. *See generally* Barocas & Selbst, *supra* note 85; Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework To Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Joshua A, Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Selbst, *supra* note 80.

[97] *See* Hylton, *supra* note 79, at 48 (noting how technological change may increase activity levels and social cost).

false negatives, but many false positives.) Even so, the data and reputational scores used and produced by the algorithm may feed into databases used by others, including not only future employers but many other decision-makers operating in many other contexts. Or imagine a policing algorithm that sends police to neighborhoods where police have already been arresting people, thus reinforcing the notion that the area is especially high crime and needs additional police surveillance.

The point of the Algorithmic Society is to increase the opportunity, speed, and cost-effectiveness of decision-making. Firms employ algorithms to save money, and to perform repetitious tasks and calculations on a vast scale that would be prohibitively expensive, or even impossible, for humans to perform. This allows firms to ask questions that would previously have been unanswerable and to make decisions that would previously have been prohibitively expensive to formulate and adopt.

This phenomenon—more kinds of decisions more cheaply made—is just another example of the substitution effect that is characteristic of robotics generally. We substitute algorithmic judges and calculators for human ones. But the choice of algorithms, the choice of categories, the kind of data collected, and the distributed creation and maintenance of digital identities have social costs, whose burden is shifted onto others—onto the general population, or onto particular segments of the population, like the poor, or minority communities.

In addressing algorithmic harms and algorithmic discrimination, our goal is not smoking out bad intention. Posing the question that way is yet another example of the homunculus fallacy. Rather, the goal, as more and more companies shift to algorithmic decision-making, and increase their levels of decision-making activity, is to require firms to adopt methods that are justified from the standpoint of society as a whole. Just as in the case of public nuisance, the state must decide how best to make businesses internalize their costs.

Frank Pasquale has also pointed out that, to the extent we use this approach to algorithmic nuisance, we must be able to identify the persons or organizations who are using the algorithm that is imposing costs on the rest of society.[98] That is, algorithms must be designed so that we can tell which persons or organizations are employing them. The Golem must be traceable to a Rabbi or group of Rabbis. In many cases—finance, employment, policing—identification of the user will not be difficult because the organization that is using the algorithm identifies itself. But in many cases, the algorithm and the data it uses will have been constructed by many organizations working together. There could also be cases in which algorithmic decision-making is made by anonymous or pseudonymous persons or organizations. Then the law will have to require disclosure of who is behind the algorithm in order to enforce a public duty.

---

[98] Frank Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1239, 1248–51 (2017).

VIII. CONCLUSION

The hope and the fear of robotics and artificial intelligence has been with us from the earliest days of literature. Even today, journalists write stories raising the specter of out-of-control robots, algorithms, and AI agents who will soon take over our world. My goal in this Lecture has been to offer a corrective. Talking this way locates the danger of the Algorithmic Society in the robots, AI agents, and algorithms themselves. But the true danger is, and always has been, in the people—the organizations and businesses that adopt and employ these devices, and use them to affect, control, and manipulate other human beings. If we were all as pious as the Maharal, we would not need to fear the Golem. Because we are not, we need to learn how to restrain ourselves.