Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society

FRANK PASQUALE*

TABLE OF CONTENTS

I. INTRODUCTION

Jack Balkin makes several contributions in this Lecture. He proposes a set of "laws of robotics" for an "Algorithmic Society" (i.e., one characterized by "social and economic decision-making by algorithms, robots, and AI agents").[1] These laws both elegantly encapsulate and add new principles to a growing movement for accountable design and deployment of algorithms.[2] My purpose in commenting on his Lecture is threefold: (1) to contextualize his proposal as a kind of regulation of regulation, familiar from the perspective of administrative law, (2) to expand the range of methodological perspectives capable of identifying "algorithmic nuisance," a key concept in this Lecture,[3] and (3) to propose a fourth law of robotics to ensure the viability of Balkin's first three laws.

Balkin argues 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."[4] In response to these

---

* Professor of Law, University of Maryland Francis King Carey School of Law; J.D., Yale Law School; MPhil, Oxford University; BA, Harvard University.

[1] Jack Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 Ohio St. L.J. 1217, 1217, 1219 (2017). Algorithmic information processing is in effect the "brain" of robotics and AI agents. *See generally* PEDRO DOMINGOS, THE MASTER ALGORITHM (2015). Balkin warns that "[t]he dream of the Algorithmic Society is the omniscient governance of society." Balkin, *supra*, at 1226.

[2] For more on this movement, see Frank Pasquale, *Digital Star Chamber*, AEON (Aug. 18, 2015), https://aeon.co/essays/judge-jury-and-executioner-the-unaccountable-algorithm [https://perma.cc/N4DA-BLNJ] ("Algorithmic accountability is a big tent project, requiring the skills of theorists and practitioners, lawyers, social scientists, journalists and others. It's an urgent, global cause with committed and mobilised experts looking for support.").

[3] Balkin, *supra* note 1, at 1232–33.

[4] While Balkin crystallizes this set of problems of algorithms in his discussion of his proposed third law, it buoys the normative justification for all three of his laws. *Id.* at 1239.

and other problems caused by algorithmic processing of data, he proposes three laws for an Algorithmic Society:

> (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.
> (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.[5]

Each of these laws will have important implications for the future of legal regulation of technology.

II. THE REGULATION OF REGULATION

I believe that Balkin's concept of information fiduciary is well developed and hard to challenge.[6] In our increasingly Algorithmic Society, software-driven devices are increasingly taking on roles once reserved to professionals with clear fiduciary duties.[7] A manufacturer of a medical device offering diagnoses should be held to the same standards we would impose on the physician it is replacing. Otherwise, the legal playing field will be unfairly tilted—holding physicians to standards that their would-be robotic replacements can evade.

Software-based replacements for attorneys should also operate on a level playing field. Sadly, federal antitrust policymakers have ignored this principle in the service of a broadly anti-labor agenda.[8] When North Carolina attempted to modernize its regulation of software-based legal services by preventing legal software manufacturers from foisting terms of service on users that denied them consumer protections afforded to clients of attorneys, the Federal Trade Commission and Department of Justice weighed in to criticize the state and threaten antitrust action against it.[9] Framed as an attack on attorney self-

---

[5] *Id.* at 1227 (footnote omitted).

[6] *See id.* at 1227–31.

[7] *Id.* at 1230–31.

[8] For a precis on this trend in antitrust enforcement, see Frank Pasquale, *When Antitrust Becomes Pro-Trust: The Digital Deformation of U.S. Competition Policy*, 2 ANTITRUST CHRON. 46, 47–48 (2017), https://www.competitionpolicyinternational.com/wp-content/uploads/2017/05/AC_May.pdf [https://perma.cc/5LX2-9V4D].

[9] *See generally* Joint Letter from Marina Lao, Dir., Office of Policy Planning, Fed. Trade Comm'n & Robert Potter, Chief, Legal Policy Section, Antitrust Div., U.S. Dep't of Justice, to Bill Cook, Senator, N.C. Senate (June 10, 2016), https://www.ftc.gov/system/files/

protection, the agencies' intervention had flimsy foundations in economic policy, and evidenced little to no awareness of literature on the pitfalls of automation.[10] They appear committed to promoting software as a substitute for attorneys, even though the sellers of such software often include exculpatory clauses (or other limitations of liability) that severely disadvantage users.[11] Such clauses prematurely extinguish litigation over bad outcomes, which could help both attorneys and consumers better understand the risks involved in AI approaches to law.[12]

Balkin's second law—imposing public duties—also has substantial support in extant literature. Lori Andrews proposed a bill of rights for users of social media platforms;[13] I have applied the idea of business affected with the public interest to search engines.[14] Balkin is right to support initiatives to bring such accountability to algorithmic systems generally—especially ones that have physical effects on individuals or the environment. Even Uber, a notoriously lawless firm, has indicated its interest in giving drivers an opportunity to contest algorithmic "deactivations" (i.e., firings).[15] This step toward due process is

---

documents/advocacy_documents/comment-federal-trade-commission-staff-antitrust-division-addressing-north-carolina-house-bill-436/160610commentncbill.pdf [https://perma.cc/X47U-3EZR].

[10] *See* Sandeep Vaheesan & Frank Pasquale, *The Politics of Professionalism: Reappraising Occupational Licensure and Competition Policy*, ANN. REV. L. & SOC. SCI. 7 (forthcoming) (manuscript at 7), https://ssrn.com/abstract=2881732 [https://perma.cc/DT88-5QJK] (arguing that while automation is supported by claims of efficiency, "[e]fficiency maximization is laden with implicit political judgments on the role of the state, the existing distribution of wealth, and human behavior").

[11] *See, e.g.*, H.R. 436, 2015 Gen. Assemb., 2015 Sess. (N.C. 2016) (exempting certain website providers from the definition of the "practice of law" and creating additional requirements for website providers).

[12] *See* MARGARET JANE RADIN, BOILERPLATE: THE FINE PRINT, VANISHING RIGHTS, AND THE RULE OF LAW 139–40 (2013) (describing suboptimal social outcomes arising out of exculpatory clauses).

[13] LORI ANDREWS, I KNOW WHO YOU ARE AND I SAW WHAT YOU DID: SOCIAL NETWORKS AND THE DEATH OF PRIVACY 189–91 (2011) (proposing a "Social Network Constitution"); *see also* REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM 179–80 (2012) (describing the Global Network Initiative, which is "dedicated to helping Internet and telecommunications companies uphold their users' and customers' rights to freedom of expression and privacy around the world in ways that are credible and accountable").

[14] *See* Frank Pasquale, *Dominant Search Engines: An Essential Cultural & Political Facility*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET 401, 401–02 (Berin Szoka & Adam Marcus eds., 2010); Frank Pasquale, *Internet Nondiscrimination Principles: Commercial Ethics for Carriers and Search Engines*, 2008 U. CHI. LEGAL F. 263, 277–79.

[15] Josh Eidelson, *Uber Found an Unlikely Friend in Organized Labor: A Company-Funded Guild for Drivers Promises Not to Strike*, BLOOMBERG BUSINESSWEEK (Oct. 27, 2016), http://www.bloomberg.com/news/articles/2016-10-27/uber-found-an-unlikely-friend-in-organized-labor [https://perma.cc/RJM7-L22C].

appropriate for a firm that exercises something like governance authority over important aspects of our transportation systems.[16]

Many forces tend to regulate our conduct, ranging from government to corporations to professionals. Advanced legal regimes also create safeguards to assure that regulation does not become too oppressive—in effect, modes of regulating that regulation. The Administrative Procedure Act is one example, counterbalancing the potentially overweening power of the executive branch with judicial monitoring and supervision.[17] While large online platforms increasingly act as de facto "courthouses," they are still constrained in the types of requirements they can require users to adopt.[18] Ancient professions like law and medicine self-regulate (under the active supervision of the state) to reduce the ability of rogue attorneys and doctors to abuse the power they will predictably have in important episodes of their clients' and patients' lives. Whenever algorithms move into any of these areas, new forms of regulation of regulation will need to complement the old. Without such modernization, regulatory arbitrage will artificially induce many forms of artificial intelligence to prematurely supplant more accountable institutions.[19]

---

[16] Rory van Loo, The Law of Consumer Markets (2016) (unpublished Ph.D. dissertation, Yale University) (on file with author) (giving many other examples of procedures at firms that resolve disputes by offering customers some opportunity to learn of the evidence against them, contest it, and receive a reasoned response to their complaint or appeal).

[17] *See, e.g.*, ANDREW F. POPPER ET AL., ADMINISTRATIVE LAW 59 (2d ed. 2010).

[18] Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REG. 547, 554 (2016) ("Corporations . . . design procedures and shape the de facto substantive rules governing the vast majority of consumer disputes.").

[19] For an early example of this type of imbalance, compare the application of FCRA to employers using old-school credit bureaus to investigate applicants with the free hand they have when deploying Google to assemble automated dossiers via name search queries. Frank Pasquale, *Rankings, Reductionism, and Responsibility*, 54 CLEV. ST. L. REV. 115, 136 (2006) ("The FCRA requires credit bureaus to permit individuals to dispute negative information on their credit reports and to give their own side of the story on reports generated for potential creditors, insurers, and employers."). No such requirements hamper the more automated Google name search queries in the United States, while Europe has attempted to remedy the discrepancy with prerogatives like the right of erasure and the right to be forgotten. *See* Frank Pasquale, *Reforming the Law of Reputation*, 47 LOY. U. CHI. L.J. 515, 516–17 (2015). For a definition of "premature disruption," see Brian Sheppard, *Incomplete Innovation and the Premature Disruption of Legal Services*, 2015 MICH. ST. L. REV. 1797, 1825–26, 1876 (2015) (describing how premature disruptions occur when "an industry has experienced a diminution in its capacity or willingness to meet demand for a core function at pre-disruption levels of quality, leading to a reduction in welfare that exceeds the benefits brought by the innovation" and applying this theory of premature disruption to legal services (footnote omitted)).

III. IDENTIFYING ALGORITHMIC NUISANCE

The third law is an important extension and clarification of extant concerns about algorithmic accountability and a critical intervention in these debates. Balkin's worries about the applicability of discrimination law to algorithmic processes are well founded. Activists have accused algorithmic processes of bias,[20] while their owners or programmers have deflected those charges by insisting no one at their firm intended to discriminate.[21] At present, the concept of "disparate impact" has permitted an uneasy peace between the two sides: responsible designers of algorithmic systems are committing to trying to avoid lopsided distributions of benefits or burdens with respect to historically recognized categories, such as race, gender, or sexual orientation. With respect to finance and credit determinations, regulation has already ensured that "intent" is not a *sine qua non* for liability.[22]

This is important, because, as Mireille Hildebrandt has argued, the "data-driven agency" common in algorithmic systems builds on "information and behaviour, not meaning and action."[23] Hildebrandt's contrast between

---

[20] *See, e.g.*, Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/XNH4-JV7K].

[21] *See* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 39 (2015) (discussing Latanya Sweeney's work); *see also* Brief of Defendant-Appellant at 21, State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (No. 2015AP157-CR); Brief of Plaintiff-Respondent at 21, State v. Loomis, 881 N.W.2d 749 (Wis. 2016) (No. 2015AP157-CR); WILLIAM DIETERICH ET AL., COMPAS RISK SCALES: DEMONSTRATING ACCURACY EQUITY AND PREDICTIVE PARITY 1 (July 2016), http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf [https://perma.cc/GLP4-AKPS] (responding to Angwin et al., *supra* note 20). The binary of "discrimination" or nondiscrimination may be ill-suited to probabilistic evaluations of harm. Abe Gong, *Ethics for Powerful Algorithms (1 of 4)*, MEDIUM (July 12, 2016), https://medium.com/@AbeGong/ethics-for-powerful-algorithms-1-of-3-a060054efd84#.35pjrg22k [https://perma.cc/U6VW-22FU].

[22] ROBINSON + YU, KNOWING THE SCORE: NEW DATA, UNDERWRITING, AND MARKETING IN THE CONSUMER CREDIT MARKETPLACE 21–22, 29 (2014), https://www.teamupturn.com/static/files/Knowing_the_Score_Oct_2014_v1_1.pdf [https://perma.cc/TK67-MD9H] (describing the use of fringe and alternative data for credit scoring). Employment law has related concepts. *See* Ifeoma Ajunwa, Assistant Professor, Univ. of D.C. Sch. of Law, Written Testimony from the U.S. Equal Employment Opportunity Commission Public Meeting on Big Data in the Workplace (Oct. 13, 2016), https://www.eeoc.gov/eeoc/meetings/10-13-16/ajunwa.cfm [https://perma.cc/AQP5-VAAK] (describing practices used by some employers that are prone to having discriminatory effects); Clint Boulton, *The Hidden Risk of Blind Trust in AI's 'Black Box,'* CIO MAG. (July 6, 2017), http://www.cio.com/article/3204114/artificial-intelligence/the-hidden-risk-of-blind-trust-in-ai-s-black-box.html [https://perma.cc/GD88-PAR5] (commenting that "AI in things like credit decisions, which might seem like an obvious area, is actually fraught with" legal land mines).

[23] Mireille Hildebrandt, *Law as Information in the Era of Data-Driven Agency*, 79 MOD. L. REV. 1, 2 (2016).

information and meaning here is critical. Meaning has been crucial to legal determinations. In criminal law, and even in many civil regulatory schemes (which might calibrate penalties based on willfulness), the meaning of an illegal action has been critical to the degree of culpability assigned to the defendant—and even to the definition of an act as illegal in itself.[24] But as the economy has grown more complex, strict liability grew as a conceptual resource to hold persons or firms accountable for damages they caused but may not have intended. Nuisance, too, has been a crucial tool for checking the negative effects of systems that do not rise to the level of tort.

However, there have now been so many documented instances of algorithmic discrimination that I believe Balkin may be going a bit too far in his statement, "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."[25] At this point, if a corporation decides to unleash an algorithm on Twitter substantially similar to Microsoft's Tay, it should know that there is a very high likelihood it will begin spewing racist and sexist cant within days.[26]

Researchers can no longer hide behind a shield of disruptive experimentalism to deflect responsibility for such interventions. The literature on AI ethics, algorithmic system ethics, and big data and ethics is vast, well established, and easily accessible. For example, however well protected algorithmic speech may be by the First Amendment, we should at least leave open the possibility that programmers of bots on Twitter would have some liability for defamation if, say, their creations foreseeably disseminated slurs about a private figure.[27] If that means keeping humans "in the loop" for the foreseeable future in such scenarios, so be it.[28] That requirement would not

---

[24] *See* JOSEPH VINING, FROM NEWTON'S SLEEP 299 (1995) ("There is no crime in any set of facts . . . unless there is a criminal state of mind."). *But see* Lawrence B. Solum, *Artificial Meaning*, 89 WASH. L. REV. 69, 86 (2014) ("[T]he pricing of airline fares by algorithms (rather than humans) already provides part of a legally binding agreement. We have no difficulty understanding these terms—even though they do not reflect the mental states of a human being.").

[25] Balkin *supra* note 1, at 1234.

[26] *See* Helena Horton, *Microsoft Deletes 'Teen Girl' AI After It Became a Hitler-Loving Sex Robot Within 24 Hours*, TELEGRAPH (Mar. 24, 2016), http://www.telegraph.co.uk/technology/2016/03/24/microsofts-teen-girl-ai-turns-into-a-hitler-loving-sex-robot-wit/ [https://perma.cc/2GT5-C9BG]; *see also* Chris Mills, *Why Are Microsoft's Chatbots All Assholes?*, BGR (July 4, 2017), http://bgr.com/2017/07/04/microsoft-chatbot-zo-quran-religion-plz-no/ [https://perma.cc/34DA-4B5R] (describing chatbot Zo's bizarre answer to a reporter's question as "a triple fail for Microsoft, because it's a completely nonsensical off-topic answer, wrong, and painfully insensitive").

[27] On the protection of computerized speech, see Stuart Benjamin, *Algorithms and Speech*, 161 U. PA. L. REV. 1445, 1447–49 (2013).

[28] On the right to a human in the loop, see Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation & Personhood*, 47 SOC. STUD. SCI. 216, 217, 231–32 (2017). Not only speakers, but also largely automated intermediaries

merely be a salutary channeling of technology through law.[29] It may also help address the putative future unemployment crises so often modeled as an inevitable consequence of technological development—rather than as a result of policymakers' neglect.[30]

To be sure, negligence and intent-based causes of action may fade over time if society accepts more autonomous algorithms in online and offline contexts. The effects-based regime of responsibility that Balkin proposes, like the shift from consent- to use-based regulation in data protection law, will become more important over time.

One critical question for such a regime is: how do we identify effects so negative that they merit regulatory intervention or liability? Balkin suggests a form of cost-benefit analysis: he compares algorithmic nuisance to "socially unjustified pollution" and urges policymakers to prevent algorithm users from "externalizing the costs [and harms] of their operations."[31] He argues that policymakers should focus regulation on reducing the negative effects of robotic methods that are not cost-benefit "justified from the standpoint of society as a whole."[32]

While I agree that algorithmic processing of information can impose undue costs on third parties, I hope that cost-benefit analysis will only be one of the methods that we can use to identify algorithmic nuisance. Cost-benefit analysis is manipulable, and can conceal as much as it reveals about important value judgments.[33] Deregulationists have recently deployed variants of cost-benefit

---

disseminating such speech, may face legal requirements of responsibility. *See, e.g.*, Melissa Eddy & Mark Scott, *Delete Hate Speech or Pay Up, Germany Tells Social Media Companies*, N.Y. TIMES (June 30, 2017), https://www.nytimes.com/2017/06/30/business/germany-facebook-google-twitter.html [https://perma.cc/SVB5-N5BU].

[29] For examples of such channeling, see generally LAURENCE H. TRIBE, CHANNELING TECHNOLOGY THROUGH LAW (1973) (describing the ways that the law "can be used to influence technological development"). For concrete examples of the proper role of communications professionals in automated information systems, see Frank Pasquale, *The Automated Public Sphere*, *in* BIG DATA, BIG BROTHER?: THE POLICIES AND POLITICS OF BIG DATA (Ann Rudinow Sætnan et al. eds., forthcoming 2018).

[30] On the role of law in forestalling future mass unemployment, see Frank Pasquale, *To Replace or Respect: Futurology as if People Mattered*, BOUNDARY2 (Jan. 20, 2015), http://www.boundary2.org/2015/01/to-replace-or-respect-futurology-as-if-people-mattered/ [https://perma.cc/37DY-DDXR].

[31] Balkin, *supra* note 1, at 1227, 1233 ("[W]e are talking about the socially unjustified use of computational capacities that externalizes costs onto innocent others.").

[32] *Id.* at 1240. This approach could be realized in many ways. *Ex post*, technology producers could be required to compensate those whom they harm. *Ex ante*, a data levy could collect revenue in order to pay actuarially predictable claims, given experience of data breaches and other negative effects of data collection.

[33] *See generally* FRANK ACKERMAN & LISA HEINZERLING, PRICELESS: ON KNOWING THE PRICE OF EVERYTHING AND THE VALUE OF NOTHING (2004) (arguing that the value of cost-benefit analysis is severely limited in the context of "priceless" subjects, such as human life and environmental protection); DOUGLAS A. KYSAR, REGULATING FROM NOWHERE: ENVIRONMENTAL LAW AND THE SEARCH FOR OBJECTIVITY (2010) (proposing that cost-

analysis to sandbag financial regulation, an algorithm-rich field where academic-industry collaborations have all too often underestimated the risks from automation and modeling, and overestimated the costs of building safeguards into them.[34]

There is also a tension between the two key metaphors in Balkin's discussion of algorithmic nuisance. He draws on environmental law,[35] a move that has served intellectual property law well,[36] as well as privacy scholars, who have analogized the diffuse harms arising out of misuse of data, to the harms finally recognized and countered by environmental law after property law failed to address them.[37] But the appeal to environmental law principles fits uneasily with the economism of cost-benefit analysis. Many forms of environmental regulation that we now consider sacrosanct (think, for example, of Bill Eskridge's characterization of the Endangered Species Act as a "super-statute"[38]) would not survive contemporary Office of Information and Regulatory Affairs (OIRA)-driven cost-benefit standards.

Of course, it is possible to enter the field of cost-benefit battles and win. But it's important to maintain deontological patterns of justification in the technology world to complement the utilitarianism of cost-benefit analysis. For example, consider a firm that deploys resume-filtering software to choose new employees who will be like the employees who succeeded in the firm in the past. This could be based on thousands of variables, including available social network data of current employees and applicants. If the firm has had diversity problems in the past, it is likely to maintain this pattern of problems with such software.[39] It is hard to quantify the cost of such lack of diversity, but as Jerry Kang has observed, it is not necessary for us to do so: such discrimination is in itself objectionable, regardless of its effects.[40]

---

benefit analysis must be overhauled so that it factors context and long-term effects into its decision-making).

[34] *See* John C. Coates IV, *Cost-Benefit Analysis of Financial Regulation: Case Studies and Implications*, 124 YALE L.J. 882, 896 (2014).

[35] As the mention of "environmental impact statements" suggests. *See* Balkin, *supra* note 1, at 1234.

[36] *See* James Boyle, *A Politics of Intellectual Property: Environmentalism for the Net?*, 47 DUKE L.J. 87, 108–12 (1997); Frank Pasquale, *Toward an Ecology of Intellectual Property: Lessons from Environmental Economics for Valuing Copyright's Commons*, 8 YALE J.L. & TECH. 78, 119–27 (2006).

[37] Dennis D. Hirsch, *Protecting the Inner Environment: What Privacy Regulation Can Learn from Environmental Law*, 41 GA. L. REV. 1, 9–10 (2006).

[38] William N. Eskridge, Jr. & John Ferejohn, *Super-Statutes*, 50 DUKE L.J. 1215, 1242–43 (2001).

[39] *See generally* CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY (2016) (arguing that machine learning based on training data from the past is not a guarantee of fairness, but rather a way of propagating past practices and automating the status quo). Moreover, thanks to activists like O'Neil's work and patient explanations in dozens of venues, the firm should at this point know that it will do so.

[40] Jerry Kang, *Race.net Neutrality*, 6 J. ON TELECOMM. & HIGH TECH. L. 1, 13 (2007).

We should also leave open the door to flat bans on certain data uses (e.g., no use of credit scores based in any way on health data), even if well-paid econometricians can demonstrate extraordinary gains in economic efficiency arising out of such classifications. Consider robotic patrols that could easily reduce injuries and increase order at protest policing. They could also raise the possibility of perfected, distant, "push button" law enforcement that has never been part of humane legal orders.[41] A just society could decide to ban such patrols, even if such a decision was most likely to decrease welfare overall, on the grounds that it was more concerned with avoiding a type of automated oppression whose likelihood of occurrence was impossible to calculate a priori.

To complement Balkin's story of the Golem,[42] I am reminded of the Gospel of Luke, where Jesus relates the parable of the lost sheep.[43] Leaving ninety-nine sheep in "open country" to find one is not necessarily an economically rational decision.[44] Similarly, the Endangered Species Act has been decried by technocrats as placing inordinate value on unusual species. But each decision reflects altogether human spontaneity and idealism. Such values should not be sacrificed in the name of some scientific monetization of the value of ecosystem services, or similar calculations.

Just as we cannot quantify in monetary terms all forms of human transformation of the natural world that are discomfiting enough to merit legal regulation, we will not always be able to offer precise valuations of the alarm or apprehension we feel at certain algorithmic transformations of human social relations. The analogy between algorithmic harm and pollution may be challenged as unduly scientific. We know that if particulate levels double, lung cancer cases will probably go up by x%. Do we know what happens to an economy if, for instance, a secret scoring mechanism consigns 1% of those scored to self-reinforcing cycles of exclusion (to continue the parable of the lost sheep)? Or if a beauty algorithm picks a disproportionately white selection of finalists?[45] We can work to avoid such harms, without trying to quantify them.

---

[41] Jathan Sadowski & Frank Pasquale, *The Spectrum of Control: A Social Theory of the Smart City*, 20 FIRST MONDAY (July 6, 2015), http://firstmonday.org/ojs/index.php/fm/article/view/5903/4660 [https://perma.cc/UKP4-3U2H]. For more on the horrors of "push button" force, see NORBERT WIENER, GOD AND GOLEM, INC.: A COMMENT ON CERTAIN POINTS WHERE CYBERNETICS IMPINGES ON RELIGION 24–25 (1964).

[42] Balkin, *supra* note 1, at 1222–23; Omer Tene & Jules Polonetsky, *Taming the Golem: Challenges of Ethical Algorithmic Decision Making*, N.C. J.L. & TECH (forthcoming 2017) (manuscript at 7–8), https://fpf.org/wp-content/uploads/2016/05/Golem_May153-1.docx [https://perma.cc/7NSK-JNN2] (examining the potential for bias and discrimination in automated algorithmic decision-making).

[43] "Suppose one of you has a hundred sheep and loses one of them. Doesn't he leave the ninety-nine in the open country and go after the lost sheep until he finds it? And when he finds it, he joyfully puts it on his shoulders and goes home." *Luke* 15:4–6.

[44] *Id.*

[45] Jordan Pearson, *Why an AI-Judged Beauty Contest Picked Nearly All White Winners*, MOTHERBOARD (Sept. 5, 2016), http://motherboard.vice.com/read/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners [https://perma.cc/4PAE-C2MR].

IV. THE ATTRIBUTION PROBLEM IN ROBOTICS

A voice-parsing algorithm might predict Supreme Court votes much more cheaply than the justices and clerks arguing and writing out decisions.[46] But no respectable legal system would substitute it for actual legal determinations, at whatever level of the justice system it might be deployed, because it cannot relate its rationale with reasons that have normative weight.[47] Explainability matters because the process of reason-giving is intrinsic to juridical determinations—not simply one modular characteristic jettisoned as anachronistic once automated prediction is sufficiently advanced.

One key element of explainability is a clear sense of the history of a robot—how was it first programmed, to what has it been exposed, and how has this interplay between hardware, software, and the external environment resulted in present behavior. At the core of Balkin's Laws of Robotics is a concern to make certain individuals (whose role parallels that of the golem-creating rabbi) responsible for their creations.[48] He does not want to create a set of legal obligations for algorithms or robots.[49] Rather, he builds on our centuries-long experience with regulating persons.[50] He observes that regulating the owners and programmers of artificial intelligence will require some monitoring of what they are creating and coding.[51] To guarantee the efficacy of such monitoring, regulators may need to establish some ground rules, or preregulation, of the interactions algorithms will have with the wider world.

I completely agree with Balkin's point that the robot as substitute for human actor all too often acts as a "fetish or deflection away from the social bases of power."[52] But there is a sizeable literature anticipating or hoping for fully autonomous robotics or software systems, unmonitored and even uncontrolled by any person.[53] This type of unmooring of machines from responsible human

---

[46] Bryce J. Dietrich et al., Emotional Arousal Predicts Voting on the U.S. Supreme Court 2–6 (Oct. 12, 2016) (draft manuscript), http://people.hmdc.harvard.edu/~renos/papers/DietrichEnosSen/DietrichEnosSen.pdf [https://perma.cc/576S-EXLU] (showing that "the higher emotional arousal, or excitement directed at an attorney compared to his or her opponent, the less likely that attorney is to win the Justice's vote" (emphasis omitted)).

[47] Frank Pasquale & Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviorism*, U. TORONTO L.J. (forthcoming 2018) (critiquing the use of machine learning and artificial intelligence to predict judicial behavior).

[48] Balkin, *supra* note 1, at 1222–23.

[49] *Id.*

[50] *Id.*

[51] *Id.* at 1223–25.

[52] *Id.* at 1224.

[53] *See generally* SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011) (discussing how our current philosophies and legal theories can accommodate the world's progressively sophisticated AI technology); Vitalik Buterin, *Cryptographic Code Obfuscation: Decentralized Autonomous Organizations Are About To Take a Huge Leap Forward*, BITCOIN MAG. (Feb. 8, 2014), https://bitcoinmagazine.com/articles/cryptographic-code-obfuscation-decentralized-autonomous-

agents is generally a terrible idea.[54] To preserve the world Balkin depicts—where there clearly is a rabbi for each golem—we may need to ensure that robots and algorithmic agents are traceable to and identified with their creators.[55]

First steps have already been taken in this direction. Analysts have proposed a "license plate for drones" to link any reckless or negligent flying to the drone's owner or controller, and drone registration now occurs in the United States.[56] Computer systems already try to solve or mitigate the "attribution problem" in cybersecurity, which occurs when someone attacks a system anonymously, by correlating signatures of action to known bad actors. There is broad moral agreement that opaque computational systems should not be allowed to make life-and-death decisions on battlefields. An international campaign to ban "killer robots" is now attempting to codify such ethical commitments.[57] And even if they do not succeed in obtaining a ban on such weapons, they should at least try to assure provenance designation on any autonomous weapon system.[58] So, too, we might propose a fourth law to complement Balkin's first three: "A robot must always indicate the identity of its creator, controller, or owner."

Such a proviso could also serve as a "zeroth" law, complementing the meta-principle that Asimov introduced as his Zeroth Law of Robotics (namely, that robots must not harm humanity).[59] In this case, the foundational status of a law of provenance arises out of its presumption that any given robot or algorithmic system has a creator, controller, or owner. The cutting edge of the AI, machine learning, and robotics fields emphasizes autonomy, whether of smart contracts, high-frequency trading algorithms (at least in time spans undetectable by humans), or future robots. There is a nebulous notion of "out of control" robots, which escape their creator's control—and even ideas that creators of such robots

---

organizations-huge-leap-forward-1391849871 [https://perma.cc/K29N-8KV9] (describing how "distributed autonomous organizations" (DAOs) run on blockchain protocols).

[54] *See generally* ILLAH REZA NOURBAKHSH, ROBOT FUTURES (2013) (imagining a future that includes robots in all aspects of life and examining the underlying technology and the social consequences of these innovations).

[55] There is already some work being done on this. *See* Joseph Lorenzo Hall, *'License Plates' for Drones?*, CDT BLOG (Mar. 8, 2013), https://cdt.org/blog/license-plates-for-drones/ [https://perma.cc/KSB7-ZU6A]; Mark Austen et al., *Requirements for a Global Legal Entity Identifier (LEI) Solution*, Trade Associations Global LEI Proposal 1, 7 (2011), http://www.gfma.org/uploadedfiles/initiatives/legal_entity_identifier_%28lei%29/requirementsforagloballeisolution.pdf [https://perma.cc/MZH7-AQTG]; OPENCORPORATES, https://opencorporates.com/info/about [https://perma.cc/5XCE-KDLB].

[56] *See* 14 C.F.R. § 48.15 (2016) (requiring the registration of all small unmanned aircraft other than model aircrafts as of August 29, 2016); Hall, *supra* note 55.

[57] *See generally* CAMPAIGN TO STOP KILLER ROBOTS, https://www.stopkillerrobots.org [https://perma.cc/N46K-9D4V] (describing a movement that encourages countries to create multilateral standards or regulations to cover AI applications).

[58] International law requires combatants to identify themselves. *See* Hague Convention Respecting the Laws and Customs of War on Land, Annex, art. 1, Oct. 18, 1907, 36 Stat. 2277. Certainly, that rule should apply *a fortiori* to the weaponry they deploy. *See generally* CAMPAIGN TO STOP KILLER ROBOTS, *supra* note 57.

[59] ISAAC ASIMOV, ROBOTS AND EMPIRE 291 (1985).

should escape responsibility once that "escape" has occurred. A requirement that any AI or robotics system has some designated party responsible for its action would help squelch such ideas.

Of course, some robots and algorithms will appreciably evolve away from the ideals programmed into them by their owners, as a result of interactions with other persons and machines (think, for instance, of advanced self-driving cars that evolve as a result of multiple influences).[60] In such cases, there may be multiple potentially responsible parties (to use a term from the Comprehensive Environmental Response, Compensation, and Liability Act, or CERCLA) for any given machine's development and eventual actions.[61] Whatever affects the evolution of such machines, the original creator should be obliged to build in certain constraints on the code's evolution to a) record influences and b) prevent bad outcomes. Once another person or entity hacks or disables those constraints, the hacker is responsible for the robot's wrongdoing.

For a concrete application of this principle, consider a chatbot that gradually learns certain patterns of dialogue from interactions on Twitter. According to some news accounts, Microsoft's AI chatbot, Tay, quickly adopted the speech patterns of a depressed Nazi sympathizer after only a few hours on Twitter.[62] Microsoft did not program that outcome—but it should have known that it was a danger of exposing a bot to a platform notorious for its poor moderation of harassment and hate speech. Moreover, to the extent the chatbot did log where the malign influences came from, it could report them to Twitter—which could, in some better version of itself, take some action to suspend or slow the flood of abuse coming from troll accounts and worse.

The cornerstone of Balkin's proposal is to create obligations of responsibility in systems that do not necessarily share the human experience of intent. To make his principles effective, regulators will need to require "responsibility-by-design" (to complement extant models of security-by-design and privacy-by-design). That may involve requiring certain hard-coded audit logs in both closed and open robotics, or licensing practices in open robotics that explicitly contemplate problematic outcomes (like business associate agreements governing data transfers in the HIPAA context).[63] Like the "legal entity identifiers" now vital to initiatives to create a consolidated audit trail in finance, such initiatives will not simply regulate robotics post hoc, but will

---

[60] Vladeck states, "Conferring 'personhood' on these machines would resolve the agency question; the machines become principals in their own right, and along with new legal status would come new legal burdens, including the burden of self-insurance." David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117, 150 (2014).

[61] 42 U.S.C. § 9607(a) (2012) (defining four categories of potentially liable parties).

[62] Horton, *supra* note 26.

[63] *See* M. Ryan Calo, *Open Robotics*, 70 MD. L. REV. 571, 583–91 (2011); Diana Marina Cooper, *The Application of a "Sufficiently and Selectively Open License" to Limit Liability and Ethical Concerns Associated with Open Robotics*, *in* ROBOT LAW 163, 164–65 (Ryan Calo et al. eds., 2016).

necessarily influence systems development by foreclosing some design options and encouraging others.[64]

V. CONCLUSION

Balkin's Lecture is a tour de force distillation of principles of algorithmic accountability, and a bold vison for entrenching them in regulatory principles. As he observes, "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."[65] They are, therefore, critically important features of our information society which demand immediately attention from regulators. High-level officials around the world need to put the development of a cogent and forceful response to these developments at the top of their agendas. Balkin's "Laws of Robotics" is an ideal place to start, both to structure that discussion at a high level and to ground it in deeply rooted legal principles.

It is rare to see a legal scholar not only work at the deepest levels of policy (in the sense of all those normative considerations that should inform legal decisions outside of the law governing the case) but also recommend in clear and precise language a coherent set of concrete recommendations that both exemplify principles of critical and social theory, and stand some chance of being adopted by current government officials. That is Balkin's achievement in The Three Laws of Robotics in the Age of Big Data. It is work to celebrate and rally around, and an auspicious launch for Ohio State's program in Big Data & Law.

---

[64] STAFF OF THE U.S. SEC. & EXCH. COMM'N & STAFF OF THE U.S. COMMODITIES FUTURES TRADING COMM'N, JOINT STUDY ON THE FEASIBILITY OF MANDATING ALGORITHMIC DESCRIPTIONS FOR DERIVATIVES 16, 16 n.77, 24 (Apr. 2011), https://www.sec.gov/news/studies/2011/719b-study.pdf [https://perma.cc/8768-C6DK].

[65] Balkin, *supra* note 1, at 1239.