Oklahoma Law Review

Volume 72 | Number 1
Symposium: Lawyering in the Age of Artificial Intelligence

2019

The Promise and Limitations of Artificial Intelligence in the Practice of Law

W. Bradley Wendel

Recommended Citation
W. Bradley Wendel, *The Promise and Limitations of Artificial Intelligence in the Practice of Law,* 72 OKLA. L. REV. 21 (2019).

THE PROMISE AND LIMITATIONS OF ARTIFICIAL INTELLIGENCE IN THE PRACTICE OF LAW

W. BRADLEY WENDEL[*]

Abstract

Artificial intelligence has demonstrated the ability to outperform humans at tasks that were previously thought to offer a decisive advantage to human intelligence. Computer technology has already changed the practice of law in many ways. Lawyers may therefore wonder whether they will soon be replaced by computers. This Article looks at that issue from another direction, beginning with the nature of law as a means to enhance the human ethical capacity for reason-giving in response to demands for accountability. Moral reason-giving reflects the mutual recognition of two agents as free and equal. The law merely enables the process of giving reasons on a much larger scale, given background conditions of disagreement and uncertainty. The core function of lawyers is to facilitate the law's practical authority, by interpreting and applying the law to give reasons that suffice to justify actions that affect the interests of others. The Article reviews the current state of research on machine ethics and the development of artificial moral agents and concludes that human technology is a long way from being able to design a computer system that can satisfy the demand for authority and accountability that is constitutive of the core function of lawyers in a liberal democratic political community.

I. Introduction

After years of hype about their potential,[1] artificial intelligence (AI) systems have recently shown themselves capable of outperforming humans at tasks at which humans have long had a decisive advantage. The story of IBM's Deep Blue beating Garry Kasparov at chess is familiar, as is the winning performance of the company's Watson technology at *Jeopardy*. A bigger deal, however, is the victory of the AlphaGo system, built by

[*] Professor of Law, Cornell Law School. I gratefully acknowledge the research funding provided by the Judge Albert Conway Memorial Fund for Legal Research, established by the William C. and Joyce C. O'Neil Charitable Trust.

[1] *See, e.g.*, RICHARD SUSSKIND & DANIEL SUSSKIND, THE FUTURE OF THE PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS (2015); RICHARD SUSSKIND, THE END OF LAWYERS? (2008).

Google's DeepMind division, over top-ranked human Go players.[2] Go is a significantly greater challenge for artificial intelligence than chess, because there are so many possible combinations of moves in the game that a computer cannot simply calculate the best strategy by brute force. The AlphaGo team first trained an AI system using examples of human expert moves, but subsequently developed AlphaGo Zero, which is based solely on reinforcement learning beginning with the rules of the game, without any human expert input.[3] AlphaGo Zero went on to post a 100-0 record in games against the original AlphaGo, which itself had beaten human master Lee Sedol, winner of eighteen international championships.[4] The same team also developed a chess-playing computer that learned the game in the same way, by trial-and-error, beginning with only the basic rules of the game. Unlike Deep Blue and other systems that used brute force to out-compute human players, the AlphaZero chess system "played like no computer ever has, intuitively and beautifully, with a romantic, attacking style."[5] Garry Kasparov wrote that the computer had developed its own style, not that of its human programmers, and it was one that "reflects the truth" about chess.[6]

Go and chess are only games, but melanoma can be a deadly serious skin cancer. In a study published in May 2018, a team of researchers from Germany, the U.S., and France demonstrated that an artificial neural network, trained on images of malignant melanomas and benign moles, could outperform expert dermatologists at making the potentially lifesaving

[2] *See, e.g.*, David Etherington, *Google's AlphaGo AI Beats the World's Best Human Go Player*, TECHCRUNCH, https://techcrunch.com/2017/05/23/googles-alphago-ai-beats-the-worlds-best-human-go-player/ (last visited May 29, 2019); David Z. Morris, *Google's Go Computer Beats Top-Ranked Human*, FORTUNE (Mar. 12, 2016), http://fortune.com/2016/03/12/googles-go-computer-vs-human/.

[3] *See* David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354, 354 (2017).

[4] *Id.* AlphaGo Zero not only learned the strategies and techniques employed by human experts, but developed non-standard, but successful strategies "beyond the scope of traditional Go knowledge." *Id.* at 357. "AlphaGo Zero rapidly progressed from entirely random moves towards a sophisticated understanding of Go concepts, including *fuseki* (opening), *tesuji* (tactics), life-and-death, *ko* (repeated board situations), *yose* (endgame), capturing races, *sente* (initiative), shape, influence and territory, all discovered from first principles." *Id.* at 358.

[5] Steven Strogatz, *One Giant Step for a Chess-Playing Machine*, N.Y. TIMES, Dec. 26, 2018, https://www.nytimes.com/2018/12/26/science/chess-artificial-intelligence.html.

[6] *Id.*

discrimination between malignant and benign skin lesions.[7] The AI system missed fewer malignant melanomas, and also had fewer false positives, misidentifying fewer benign moles as malignant. The computer's performance on the images alone bettered that of the dermatologists even after the human physicians were given clinical information about the patient, including age, sex, and location of the lesion.[8] The authors of the study suggested that dermatologists may benefit from the assistance of the AI system,[9] but it is not difficult to imagine the consternation of human physicians at discovering that computers do better at one of their central professional tasks.

Lawyers similarly may worry about being displaced by AI systems. Many of the tasks traditionally performed by lawyers involve dealing with large volumes of information, and computers are very good at executing instructions for the processing of information.[10] Discovery practice, particularly reviewing for privileged documents, work product, and "hot docs," has been revolutionized by predictive coding systems.[11] AI systems can provide similar advantages for the transactional due diligence process—another traditional bane of the existence of junior associates. Much technology used by law firms today is used to automate "search-and-find type tasks."[12] In addition, however, it is already possible to automate the production of routine legal instruments such as wills and residential real estate closing documents. Contract drafting software supported by machine learning and deep learning techniques may enable parties to create more sophisticated contracts without the assistance of lawyers, based on the input

[7] *See* H.A. Haenssle et al., *Man Against Machine: Diagnostic Performance of a Deep Learning Convolutional Neural Network for Dermoscopic Melanoma Recognition in Comparison to 58 Dermatologists*, 29 ANNALS OF ONCOLOGY 1836 (2018).

[8] *Id.* at 1839.

[9] *Id.* at 1841.

[10] *See* John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transfer the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041 (2014); Dana Remus & Frank Levy, *Can Robots Be Lawyers?* Computers, Lawyers, and the Practice of Law, 30 GEO. J. LEGAL ETHICS 501, 508 (2017).

[11] Dana A. Remus, *The Uncertain Promise of Predictive Coding*, 99 IOWA L. REV. 1691, 1701-05 (2014). The California State Bar ethics committee has stated that the attorney's baseline duty of competence in litigation representation requires familiarity with e-discovery and may, on a case-by-case basis, require higher levels of technical knowledge and ability. *See* Cal. State Bar Standing Comm. on Prof'l Responsibility & Conduct, Formal Op. 2015-193 (2015).

[12] Steve Lohr, *A.I. Is Doing Legal Work. But It Won't Replace Lawyers, Yet*, N.Y. TIMES (Mar. 19, 2017), https://www.nytimes.com/2017/03/19/technology/lawyers-artificial-intelligence.html.

of a few key terms and conditions.[13] Other tedious tasks, like periodic reviews by banks of commercial loan agreements, can be automated with considerable savings in costs to clients, but with a corresponding loss of jobs by lawyers. For example, JP Morgan Chase deployed an AI-based program that performs a periodic review of loan agreements in seconds—a task which previously required 360,000 hours of work per year by lawyers and loan officers.[14] Decision-support systems employing predictive analytics, like LexMachina and Ravel, can help litigators and their clients make better strategic decisions by ferreting out subtle patterns in judicial decisions and predicting the odds of success on motions at various stages of the process.[15] AI-enabled systems also have significant potential to improve access to the legal system for poor and middle-income clients. An app called DoNotPay, which allows drivers to contest parking tickets without the intervention of a human agent, has been touted as a harbinger of the transformation of self-help legal services.[16] Its developer is experimenting with the use of the platform to enable applications for emergency housing assistance and even asylum in the United States or Canada for refugees.[17]

It is important not to overstate the potentially disruptive impact of artificial intelligence on the practice of law. Different types of legal work vary in their susceptibility to replacement by automation.[18] At least for the foreseeable future it seems extremely unlikely that computers will replace lawyers at tasks such as (1) fact investigation, including making a judgment about the relevant avenues of investigation, determining where relevant

[13] *See, e.g.*, Beverly Rich, *How AI Is Changing Contracts*, HARV. BUS. REV. (Feb. 12, 2018), https://hbr.org/2018/02/how-ai-is-changing-contracts.

[14] Hugh Son, *JPMorgan Software Does in Seconds What Took Lawyers 360,000 Hours*, BLOOMBERG (Feb. 27, 2017, 6:31 PM CST), https://www.bloomberg.com/news/articles/2017-02-28/jpmorgan-marshals-an-army-of-developers-to-automate-high-finance.
An Israeli startup company called LawGeex is marketing contracts-review software to businesses that performs functions similar to the program used by JP Morgan Chase. *See* Steve O'Hear, *LawGeex Raises $12M for Its AI-Powered Contract Review Technology*, TECHCRUNCH, https://techcrunch.com/2018/04/17/lawgeex-raises-12m-for-its-ai-powered-contract-review-technology/ (last visited May 14, 2019).

[15] *See* Jason Koebler, *Rise of the Robolawyers*, ATLANTIC (Apr. 2017), https://www.theatlantic.com/magazine/archive/2017/04/rise-of-the-robolawyers/517794/; McGinnis & Pearce, *supra* note 10, at 3052-53.

[16] *See* Drew Simshaw, *Ethical Issues in Robo-Lawyering: The Need for Guidance on Developing and Using Artificial Intelligence in the Practice of Law*, 70 HASTINGS L.J. 173, 174-75 (2018) (describing development and expansion of DoNotPay).

[17] *Id.* at 176.

[18] *See* Frank Pasquale & Glyn Cashwell, *Four Futures of Legal Automation*, 63 UCLA L. REV. DISC. 26 (2015).

documents are likely to be located, and interviewing witnesses;[19] (2) negotiation over issues such as the terms on which to settle a case or provisions in a transaction;[20] (3) the type of client counseling requiring emotional intelligence, such as listening empathetically to a client in a matrimonial dispute to determine the client's goals, and then providing advice about what options are legally available;[21] (4) creative, strategic advising or that which requires assessing not only legal risks but also taking into account multifaceted, ambiguous, possibly conflicting client objectives and interests; (5) producing written work product that does not read like it was written by a computer;[22] (6) in-court appearances on behalf of clients, at a trial, evidentiary hearing, or oral argument on a motion or an appeal; or (7) any work in new or rapidly changing areas of law.[23] Technology enhances the ability of lawyers to deal with large volumes of information and to discern patterns in what may initially appear to be attributed to randomness, but experienced lawyers are still required to interact with many types of clients and exercise judgment on their behalf. Risk-averse clients may rely on automation for relatively routine legal services, but will still prefer human lawyers for high-stakes legal matters.[24] One Silicon Valley lawyer interviewed in the *New York Times* noted that the kind of work appropriate for a senior partner billing $1200 per hour is not threatened by artificial intelligence, but he did observe the potential impact on the work generally performed by more junior lawyers. He said: "For the time being, experience like mine is something people are willing to pay for . . . . What clients don't want to pay for is any routine work. But . . . the

[19] Remus & Levy, *supra* note 10, at 527.

[20] *Id.* at 527-29. Remus and Levy discuss a company called Modria that offers technology for handling relatively small disputes. It uses software that identifies "areas of agreement and disagreement, and makes suggestions for resolving the dispute." *Id.* at 528. There is a significant gap, however, between the abilities of Modria's current system and what would be required to negotiate larger, more complex disputes.

[21] *See* AUSTIN SARAT & WILLIAM L. F. FELSTINER, DIVORCE LAWYERS AND THEIR CLIENTS: POWER AND MEANING IN THE LEGAL PROCESS 53-63 (1995) (describing the fluidity of client goals, objectives, and expectations, and the counseling required to match up the client's conception of his or her interest with what is "realistic" or legally possible).

[22] *See* Lohr, *supra* note 12 (describing a program from Ross Intelligence, that replies to a legal question in the form of a two-page memo but noting that humans must rewrite the computer-generated memo).

[23] *See* McGinnis & Pearce, *supra* note 10, at 3042.

[24] Pasquale & Cashwell, *supra* note 18, at 40 (noting that "risk aversion may trump technology diffusion").

trouble is that technology makes more and more work routine."[25] As routine legal services are commodified and automated, lawyers whose bread-and-butter work consists predominantly of routine tasks such as simple wills and residential real estate closings will come under considerable competitive pressure.

One might of course argue that technology has not yet developed to handle the more complex challenges handled by the $1200 per hour partner. Given the history of advances in AI and the increase in computing power that we have seen over the last decade, however, it would be foolish to assume that the necessary technology will not exist soon.[26] In this paper I would like to make a very different claim—one grounded in moral philosophy and reflection on the nature of law. My claim is that computers can never displace lawyers entirely, because legal reasoning necessarily involves the types of normative judgments that are impossible for AI. The reason is not related to current or foreseeable limitations in computing power or AI technology. It is related instead to a conceptual truth about law, namely that it purports to impose obligations or confer rights, to give reasons, and to change what its subjects ought to do.[27] By its nature, the law claims authority.[28] Having authority means altering the normative situation of a subject; it means possessing the power to change what someone else ought to do.[29] A judge announcing a legal decision is not merely saying that the parties to whom it applies will be subject to contempt of court penalties if they disobey it (although the decision does imply the possibility of sanctions).[30] Beyond that, the judge's decision creates an *obligation* which, as H.L.A. Hart famously argued, is different from the compulsion to do

[25] Lohr, *supra* note 12 (quoting James Yoon, a partner at Wilson Sonsini Goodrich & Rosati). John McGinnis and Russ Pearce predicted this result when they wrote in 2014 that "superstars in the profession will be more identifiable and will use technology to extend their reach." McGinnis & Pearce, *supra* note 10, at 3042.

[26] *See* McGinnis & Pearce, *supra* note 10, at 3043-44 (discussing the continuing validity of Moore's Law, which predicts that computing power will continue to double every 18 months, and also noting the growth of communications bandwidth and storage capacity).

[27] *See* SCOTT J. SHAPIRO, LEGALITY 181-82 (2011); JOSEPH RAZ, THE AUTHORITY OF LAW 29-32 (1979) [hereinafter RAZ, AUTHORITY OF LAW].

[28] *See generally* JOSEPH RAZ, *Authority, Law, and Morality*, *in* ETHICS IN THE PUBLIC DOMAIN 210 (1994) [hereinafter RAZ, *Authority, Law, and Morality*].

[29] SHAPIRO, *supra* note 27, at 182; JOSEPH RAZ, THE MORALITY OF FREEDOM 26-30 (1986) [hereinafter RAZ, MORALITY OF FREEDOM].

[30] *See generally* FREDERICK SCHAUER, THE FORCE OF LAW (2015) (arguing that Hart and other canonical legal philosophers downplay the centrality of coercion in conceptual accounts of the nature of law).

something.[31] An obligation implies the existence of a duty, i.e., something which one has a good reason of the right sort to do.[32] For a reason to be of the right sort, and to affect the normative situation of those to whom it is addressed, it must express recognition of its addressee as a rational being capable of understanding and acknowledging the force of the reason.[33] Morality flows from the *mutual* recognition of one another as free and equal agents.[34] There are computational challenges involved in modeling moral decision-making by human agents.[35] Even if these problems were solved, however, it is far from clear that it is possible to produce an artificial system that counts as a good moral agent.[36] Without a moral agent, there can be no law—that is the conceptual claim I will defend here. Before elaborating on that argument, however, it is necessary to review briefly the state of the art in the field of artificial, computer, or machine ethics to see whether an AI system can be a moral agent.[37]

There are a number of ethical—in the sense of law-of-lawyering or professional responsibility—issues relating to artificial intelligence in the practice of law, which are not addressed here. These include the scope of malpractice liability for technology-assisted legal services,[38] whether legal documents or advice generated by artificial intelligence constitute the practice of law for the purpose of state prohibitions on the unauthorized

[31] H.L.A. HART, THE CONCEPT OF LAW 19, 82-88 (2d ed. 1994).

[32] STEPHEN DARWALL, THE SECOND-PERSON STANDPOINT: MORALITY, RESPECT, AND ACCOUNTABILITY 15-16 (2006).

[33] Jeremy Waldron, *The Concept and the Rule of Law*, 43 GA. L. REV. 1, 26-27 (2008) (arguing that legal authorities necessarily appeal "to people's capacities for practical understanding, for self-control, and for the self-monitoring and modulation of their own behavior, in relation to norms that they can grasp and understand").

[34] DARWALL, *supra* note 32. *See generally* CHRISTINE M. KORSGAARD, *The Reasons We Can Share: An Attack on the Distinction Between Agent-Relative and Agent-Neutral Values*, *in* CREATING THE KINGDOM OF ENDS 275 (1996); T.M. SCANLON, WHAT WE OWE TO EACH OTHER (1998).

[35] *See* Wendell Wallach, *Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making*, 12 ETHICS & INFO. TECH. 243, 244–45 (2010).

[36] *See* Colin Allen et al., *Prolegomena to Any Future Artificial Moral Agent*, 12 J. EXPERIMENTAL & THEORETICAL ARTIFICIAL INTELLIGENCE 251 (2000) [hereinafter Allen et al., *Prolegomena*]; James H. Moor, *The Nature, Importance, and Difficulty of Machine Ethics*, IEEE INTELLIGENT SYS., July/Aug. 2006, at 18, 21.

[37] Following standard usage in philosophy, I use "ethics" and "morality" interchangeably. *See* David Copp, *Introduction: Metaethics and Normative Ethics*, *in* THE OXFORD HANDBOOK OF ETHICAL THEORY 3, 4 (David Copp ed., 2006).

[38] *See, e.g.*, Benjamin H. Barton, *Some Early Thoughts on Liability Standards for Online Providers of Legal Services*, 44 HOFSTRA L. REV. 541, 557–58 (2015).

practice of law,[39] and the requirements of competence, confidentiality, and supervision when lawyers supervise non-lawyer information technology providers.[40] Lawyers concerned with managing their exposure to discipline and civil liability should obviously be well informed on the applicable law. The question addressed here is different in that it deals with whether it would be possible to create an AI system—a robo-lawyer if you will—that is capable of dealing with the law in the same way a human lawyer would.

In the discussion that follows, I will assume that the current state of AI technology permits computer systems to perform many lawyering tasks at a level of competency that meets or exceeds human lawyers. Reviewing tens of thousands of pages of documents for privileged communications, scanning hundreds of contracts for relevant provisions, generating legal documents in response to user input, and assessing the decided cases in a jurisdiction to determine the likelihood of prevailing on a motion are all functions that human lawyers have traditionally performed, but computers may do better. The claim in this paper relates to what I will call the *core lawyering function*—that which makes the legal profession ethically distinctive in the context of a liberal democracy. The core lawyering function is facilitating clients' capacity to function as free and equal members of a political community.[41] In a liberal democracy, the legal system provides individuals and entities with a toolkit of rights and duties with respect to each other, which serve as a means of treating other members of the political community with respect, in light of their inherent dignity. The law functions by offering justifications for actions that affect the interests of others. In advising clients and representing them in dealing with others, lawyers must be prepared to offer arguments that can be assessed for their soundness and accepted as reason-giving by other rational agents. The linchpin is the idea of practical authority. In ethics, that means addressing a request to another to either do something or justify her refusal

[39] *See, e.g.*, Janson v. LegalZoom.com, Inc., 802 F. Supp. 2d 1053, 1057 (W.D. Mo. 2011); Susan Saab Fortney, *Online Legal Document Providers and the Public Interest: Using a Certification Approach to Balance Access to Justice and Public Protection*, 72 OKLA. L. REV. 91 (2019); Thomas E. Spahn, *Is Your Artificial Intelligence Guilty of the Unauthorized Practice of Law?*, 24 RICH. J.L. & TECH., no. 4, 2018, at 1, 28–47.

[40] *See, e.g.*, ABA Comm'n on Ethics & Prof'l Responsibility, Formal Op. 08-451 (2008) (discussing duties regarding confidentiality when outsourcing work to IT vendors); MODEL RULES OF PROF'L CONDUCT r. 5.3 (AM. BAR ASS'N 2018) (setting out duties regarding supervision of non-lawyers assisting the practice of law).

[41] *See* W. Bradley Wendel, *The Rule of Law and Legal-Process Reasons in Attorney Advising*, 99 B.U. L. REV. 107, 109 (2018).

to do so, in response to a demand for accountability.[42] The practical authority of law comes from the claim by the legal system to create standards that must be adhered to by members of the political community because they stand for a judgment, in the name of the community as a whole, regarding what should be done in some respect.[43] The core lawyering function is precisely the connection between legal authority and the moral demand for accountability.

The reason to focus on the core lawyering function is that it clarifies the stakes in the normative debate over artificial intelligence and the law. We should be clear on what would be lost—in terms of *values*, not jobs or economic returns—by replacing human lawyers with computers or robots.

II. The Current State of Machine Ethics

As in the case of AI more generally, expert systems that participate with humans in activities with ethical significance have proven the capacity to *model* moral decision-making. For example, a team consisting of a bioethicist and a computer scientist designed an expert system called MedEthEx to help physicians deal with ethical issues that arise in the course of clinical practice.[44] The program generalizes from specific cases involving decisions made by expert human bioethicists. It uses inductive logic to infer a set of consistent rules underlying the judgments human experts have reached concerning specific cases. The system is built around the assumption that ethical decision-making proceeds from a set of principles or duties. An influential example is David Ross's list of *prima facie* duties, including fidelity, justice, gratitude, beneficence, non-maleficence, and self-improvement.[45] In the context of bioethics, the relevant principles may be those developed by Beauchamp and Childress (non-maleficence, beneficence, autonomy, and justice), and widely considered the dominant theoretical framework.[46] The expert system modeling may infer from a physician's decision to defer to her patient's refusal of life-saving treatment that respect for the patient's autonomy

[42] *See* DARWALL, *supra* note 32, at 58-59.

[43] *See* RAZ, AUTHORITY OF LAW, *supra* note 27, at 51-52.

[44] *See* WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG 127-29 (2009) (describing MedEthEx).

[45] *See* W. D. ROSS, THE RIGHT AND THE GOOD 21 (1930).

[46] *See generally* TOM L. BEAUCHAMP & JAMES F. CHILDRESS, PRINCIPLES OF BIOMEDICAL ETHICS (7th ed. 2012). For the indebtedness of Beauchamp and Childress to Ross, see Tom L. Beauchamp, *Principlism and Its Alleged Competitors*, 5 KENNEDY INST. OF ETHICS J. 181, 183 (1995).

outweighs the obligation of beneficence in the relevant circumstances. When the system is "trained" on a sufficient number of previously decided cases, it develops the capability to track human judgments with a high degree of reliability. It therefore offers an accumulated store of experience to practitioners who may not have access to advice by human bioethicists.

This short description of the MedEthEx system suggests both the promise and the limitations of AI in ethical decision-making. First, it is important to see that the system begins with a training set of cases labeled with human judgments indicating which decision is morally correct.[47] This inductive approach is an alternative to using abstract rules for the selection of ethically appropriate outcomes, reasoning from the top down.[48] Candidates for these general rules include the Ten Commandments, the Kantian Categorical Imperative, or the utilitarian calculus, and of course the first question that occurs to many people thinking about machine ethics is whether an AI system should be oriented toward a particular religious or philosophical conception of morality.[49] Beyond the obvious problem of specifying the content of morality at the most general level, the top-down approach runs into a number of significant difficulties. The decision in a particular case may involve a conflict between two or more rules on a list. These conflicts are "computationally intractable" without some higher-order rule or principle that resolves the conflict.[50] There may also be instances in which violating a rule is permissible, but again some additional principle is required to determine when exceptions are to be allowed.[51] Even in the absence of conflicts among rules, a decision-making procedure

[47] *See* Vincent Conitzer et al., *Moral Decision Making Frameworks for Artificial Intelligence*, *in* PROCEEDINGS OF THE THIRTY-FIRST AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE 4831 (Satinder P. Singh & Shaul Markovitch eds., 2017), https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14651/13991.

[48] *See* Colin Allen et al., *Artificial Morality: Top-Down, Bottom-Up, and Hybrid Approaches*, 7 ETHICS & INFO. TECH. 149, 150 (2005) [hereinafter Allen et al., *Artificial Morality*].

[49] *See* Wendell Wallach et al., *Machine Morality: Bottom-Up and Top-Down Approaches for Modeling Human Moral Faculties*, 22 AI & SOC'Y 565, 567 (2008) ("Should the systems' decisions and actions conform to religiously or philosophically inspired value systems, such as Christian, Buddhist, utilitarian, or other sources of social norms? The kind of morality people wish to implement will suggest radical differences in the underlying structure of the systems in which computer scientists implement that morality.").

[50] Allen et al., *Artificial Morality*, *supra* note 48, at 150. Faced with the problem of conflicting *prima facie* duties, Ross quotes Aristotle as maintaining "the decision rests with perception" and states that the agent's action should be "preceded and informed by the fullest reflection we can bestow on the act in all its bearings." ROSS, *supra* note 45, at 42.

[51] Allen et al., *Prolegomena*, *supra* note 36, at 254.

or system must face application questions. For example, to apply the principle "do no harm," one must be able to specify what counts as a harm. A harm may be defined as a setback to one's interests, but this merely pushes the question back another level: What interests are others ethically obligated to avoid interfering with? Not every desire rises to the level of an interest; not every unpleasant occurrence is a harm; and not every setback to one's interests is unjustified.[52] A great deal of moral discernment must go into determining the application of the seemingly straightforward harm principle. Similarly, the utilitarian calculus runs into familiar difficulties. It directs agents to lead to the greater realization (or maximize, or satisfice) the good, but the good must be specified as hedonic pleasure; higher or lower pleasures (Bentham's famous line about push-pin being as good as poetry); satisfaction of preferences (whether actual preferences or those that are fully informed or well considered); or something objective and independent of contingent psychological desires. One must further consider whether it is the good for humans or all sentient creatures, whether some things are good apart from the contribution they make to the lives of human beings, and whether it is permissible to limit the obligation to promote only the well-being of those creatures the agent has in her power to affect.[53]

More subtle ethical principles conceal even more difficult challenges. Suppose an AI system is programmed to regard violations of the Categorical Imperative as moral wrongs. On one version (the Formula of Universal Law), the morally right action is that which is according to a maxim (or principle) that one can will that it should be a universal law.[54] It is by now a standard observation that acts may fall under many descriptions,[55] so what is the maxim that is subjected to the test of universalizability? One answer—though by no means the only one—is that the proper act-description is given by what the agent intended by the action, which is consistent with Kant's usage of the term "maxim."[56] But then how

[52] *See generally* JOEL FEINBERG, HARM TO OTHERS: THE MORAL LIMITS OF THE CRIMINAL LAW 31-51 (1984) (considering the relationship among harms, wrongs, and interests).

[53] *See, e.g.*, David O. Brink, *Some Forms and Limits of Consequentialism*, *in* THE OXFORD HANDBOOK OF ETHICAL THEORY, *supra* note 37, at 381, 381–83; WALLACH & ALLEN, *supra* note 44, at 87-89.

[54] IMMANUEL KANT, GROUNDING FOR THE METAPHYSICS OF MORALS *421 (James W. Ellington trans., Hackett Publishing 2d ed., 1981) (1785).

[55] *See, e.g.*, G. E. M. Anscombe, *Modern Moral Philosophy*, 33 PHILOSOPHY 1 (1958).

[56] *See* ONORA O'NEILL, ACTING ON PRINCIPLE: AN ESSAY ON KANTIAN ETHICS 13-15 (2d ed. 2013).

would a computer have access to the motives behind any given action,[57] even assuming that people were sincere or reliable reporters of their own motives? In addition, a problem familiar to Kant scholars is determining when a maxim cannot be generalized without contradiction. Is the test one of logical contradiction or a practical one, looking at whether actions would frustrate their own purpose if performed?[58] Kant's example of a maxim that cannot be generalized as a universal law is breaking a promise to repay a debt; the idea is that if everyone did this, the practice of lending money would break down because no one would believe in another's promise to repay.[59] Returning to the earlier observation, however, acts cannot be generalized, only the principles underlying them, and if any given act exemplifies numerous principles, what should be the computer test for universalizability?[60] This is a very tough nut to crack for moral philosophers, and would appear to be an obstacle to modeling compliance with the Categorical Imperative.

The theoretical and computational challenges seem insurmountable for a top-down, theory-driven approach to machine ethics. At the very least, careful reflection, and probably also training in ethical theory, would be required to specify the goal for the system.[61] Bottom-up approaches therefore seem promising.[62] Bottom-up design in the context of artificial ethical systems means that the designer does not begin with an explicit ethical theory, bypassing many of the questions just raised. What must be specified, however, is some kind of performance measure so that engineers can tinker with the system to approach or exceed the required benchmark.[63] The system may then be turned loose in a structured environment populated by other entities with whom the system may interact. Experiments have been conducted with iterated prisoner's dilemma games or similar environments in which the evolving behavior of artificial systems can be observed.[64] Some machine ethicists have proposed a Moral Turing Test, in

[57] Allen et al., *Artificial Morality*, *supra* note 48, at 150.

[58] *See* CHRISTINE KORSGAARD, *Kant's Formula of Universal Law*, *in* CREATING THE KINGDOM OF ENDS, *supra* note 34, at 77, 77–78.

[59] KANT, *supra* note 54, at *422.

[60] O'NEILL, *supra* note 56, at 60-61.

[61] *See* Allen et al., *Artificial Morality*, *supra* note 48, at 150, 152.

[62] *See* Susan Leigh Anderson, *Asimov's "Three Laws of Robotics" and Machine Metaethics*, 22 AI & SOC'Y 477, 482 (2008) (citing W.D. Ross and the problem of conflicts of duties, but then arguing that it is possible "that a decision procedure could be learned from generalizing from intuitions about correct answers in particular cases").

[63] Wallach et al., *supra* note 49, at 569.

[64] WALLACH & ALLEN, *supra* note 44, at 101-04.

which humans have conversations, presumably in writing, about ethics with either a human or a machine, and must identify their interlocutor "at a [level of accuracy] above chance."[65] A bottom-up approach, evaluated using a Moral Turing Test, might "treat[] normative values as being implicit in the activity of agents rather than explicitly articulated (or even articulable) in terms of a general theory."[66] But this would be a rather undemanding version of the Turing Test. As the test's originators rightly note, one might also expect one's interlocutor to be able to *articulate* moral judgments and the reasons underlying them.[67] An ethical action must be justified by sufficient grounds, and in ethical discourse one might expect to be required to make explicit the reasons for one's action and their adequacy to warrant the belief that the action is morally justified. Thus, a Moral Turing Test should require demonstrated competence in explaining the reasons for a decision or action, not the action alone.

One difficulty with the Moral Turing Test is that it does not avoid contested issues in normative ethics and metaethics. For example, Mill believed that motivation was irrelevant to the morality of actions, and "[h]e who saves a fellow creature from drowning does what is morally right, whether his motive be duty or the hope of being paid for his trouble."[68] Kant, on the other hand, argued that only acts undertaken for the sake of duty have moral worth. In his shopkeeper example, Kant argued that charging a fair price does not have moral worth because it was done out of the shopkeeper's self-interest in not acquiring a reputation for dishonesty.[69] In normative ethics, a committed believer in the rights of non-human animals (Peter Singer, for example) would insist that the suffering of all sentient creatures be taken into account in a utilitarian calculus, while other philosophers would count only the pleasure and pain of humans. One response to this difficulty would be to stipulate as part of the Moral Turing Test that disagreement about either the justification of ethical judgments or the content of those judgments is consistent with competence in ethical

[65] Allen et al., *Prolegomena*, *supra* note 36, at 254.

[66] Wallach et al., *supra* note 49, at 569.

[67] Allen et al., *Prolegomena*, *supra* note 36, at 254.

[68] JOHN STUART MILL, UTILITARIANISM 24 (Oskar Priest ed., Liberal Arts Press 1957) (1863).

[69] KANT, *supra* note 54, at *397. The idea is that moral worth is connected with rational necessity, and the connection between the shopkeeper's actions and his reasons for acting is contingent on the overlap between honesty and his professional success. *See also* Jens Timmermann, *Acting from Duty: Inclination, Reason and Moral Worth*, *in* KANT'S GROUNDWORK FOR THE METAPHYSICS OF MORALS: A CRITICAL GUIDE 45, 47-48 (Jens Timmermann ed., 2009).

decision-making. If we do not expect agreement among human decisionmakers—even those with training in moral philosophy—than it is unrealistic to expect more from a machine.

In engineering terms, a bottom-up approach to problem-solving contemplates the iterated development of discrete subsystems that work together to accomplish the specified goal, such as passing the Moral Turing Test. As three leading machine ethicists have observed, computer scientists are working on subsystems that are capable of modeling particular skills and capacities that are relevant to moral decision-making.[70] However, the task of designing an integrated *artificial* moral decision-making system is complicated by our still rapidly developing understanding of how *human* moral decision-making actually works. For example, picking out the morally salient features of a situation is a complex capacity that may proceed largely unconsciously, and is probably supported by emotional intelligence as well as theoretical reasoning.[71] (Hume, of course, made emotions such as sympathy and trust central to his ethical theory.[72]) Jonathan Haidt's work has shown that many moral judgments begin with intuitions—relatively fast, automatic, affective responses—and are supported only after the fact with reasoning that is backfilled to fit the judgment already reached on the basis of unconscious factors.[73] Emotions also provide important channels for acquiring information relevant to moral decision-making.[74] The human ability to intuit the affective states of others—whether it is pain, fear, surprise, humiliation, anger, disgust, or another emotion—is essential to our ability to direct one's actions in an appropriate way. Emotional intelligence is also important in assessing the intentions of others, and if moral evaluation requires taking intentions into account (as it may on a Kantian approach), then a system that is fully competent as an artificial moral agent would require the capacity to interpret behavior in light of intentions.[75] The field of affective computing is in its infancy,[76] and significant progress in this discipline may be required

[70] Wallach et al., *supra* note 49, at 570.

[71] *See* Wendell Wallach, *Implementing Moral Decision Making Faculties in Computers and Robots*, 22 AI & SOC'Y 463, 469 (2008).

[72] *See* Annette Baier, *Hume's Place in the History of Ethics*, *in* THE OXFORD HANDBOOK OF THE HISTORY OF ETHICS 399 (Roger Crisp ed., 2013).

[73] Jonathan Haidt, *The New Synthesis in Moral Psychology*, 316 SCIENCE 998, 998 (2007); *see also* Jonathan Haidt, *The Emotional Dog and Its Rational Tail*, 108 PSYCHOL. REV. 814 (2001).

[74] WALLACH & ALLEN, *supra* note 44, at 140-41.

[75] *Id.* at 141.

[76] *Id.* at 152-53.

before an AI system can attain the competency required to be an artificial moral agent. It is likely that a bottom-up approach to machine ethics will involve subsystems that are capable of dealing with the affective as well as the cognitive demands of moral decision-making.

III. The Relationship Between Law and Morality and What It Has to Do with AI

One who believes in the potential of artificial intelligence to take over the core lawyering function of advising and representing clients within the law has several possible responses to the current state of AI moral decision-making. The first is to deny the need for moral decision-making in connection with the interpretation and application of law. As discussed in Section III.A, below, that is a mischaracterization of legal positivism, to say nothing of anti-positivism. No matter what theory about the nature of law to which one subscribes, moral decision-making is inevitable. The second, and a weaker position, would be to assert that AI systems may be able to *predict* human ethical judgments, even if they are not yet capable of *making* them. That is a more challenging argument and will be taken up in Section III.B.

A. Is Moral Decision-making Part of Law?

The legal positivist has a thesis about the relationship between law and policy considerations: they are separable.[77] A norm may be part of a system called “law” without satisfying a demand that it be just, efficient, wise, or in conformity with the requirements of morality. A different way to say the same thing is to insist that only social facts can count in favor of a conclusion that some norm is law. As John Gardner puts it, it must be possible to recognize law based “on its sources, not its merits.”[78] On one version of positivism, expounded by H.L.A. Hart, law is defined as the union of primary and secondary rules.[79] Primary rules are directed at citizens and purport to permit, prohibit, or regulate conduct.[80] Criminal prohibitions are an obvious example of primary rules, but there are myriad other types, including rules of civil liability, statutes and administrative regulations concerning matters such as workplace discrimination and food safety, and legal norms that provide a toolkit for private ordering, such as

[77] *See, e.g.*, JULES COLEMAN, *Negative and Positive Positivism, in* MARKETS, MORALS AND THE LAW 3, 5 (1988).

[78] John Gardner, *Legal Positivism: 5 1/2 Myths*, 46 AM. J. JURIS. 199, 199 (2001).

[79] HART, *supra* note 31, at 94.

[80] *See id.*

rules governing contracts, wills and trusts, and the formation and governance of corporations. Secondary rules are second-order "rules about rules," establishing norm-governed patterns of creating or changing laws, adjudicating disputes that arise under primary rules and, most importantly, sorting out which norms are a part of the law and which belong solely to some other normative domain, such as morality, custom, etiquette, or the rules of some non-legal institution such as a club or a university faculty. Hart refers to this latter type of secondary rule as a "rule of recognition," and it is the linchpin in his theory.[81] A rule of recognition specifies some feature or property of another rule by which it is shown to be legally authoritative.[82] (Dworkin calls these characteristics the "pedigree" of the rule.[83]) In the United States, for example, if a text is passed by both Houses of Congress and signed by the President, the proposition it states becomes a source of law.

Legal positivism appears to avoid the problems discussed above, concerning the ability of AI systems to make (or model) moral judgments. If legal validity is a property of the behavior of judges, legislators, and other legal officials, nothing from the domain of morality need be incorporated into legal decision-making. However, this tidy story is complicated by a couple of considerations. First, there is no reason to believe that positivism must be *exclusive*—that is, that it cannot and does not incorporate moral considerations into legal decision-making. Second, even exclusive positivism requires lawyers and judges to make judgments that may be difficult for AI systems to make or model, for many of the same reasons that moral judgments are challenging for computers.

Inclusive positivists contend that moral criteria *can* figure in tests of legal validity, as long as the relevant social practices assign them that role.[84] Moral evaluation can therefore play a role in determining what the law is, as long as there is a conventional practice among the relevant officials (such as judges) of referring to moral criteria in making decisions about what the law is. Legal standards that have cognates in moral analysis, such as

[81] *Id.*

[82] *Id.* at 94-95.

[83] RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 40 (1977) (from chapter 2, "The Model of Rules I").

[84] For this phrasing of the definition of inclusive positivism, see Scott Hershovitz, *The End of Jurisprudence*, 124 YALE L.J. 1160, 1166 (2015). *See also* Jules Coleman, *Authority and Reason*, *in* THE AUTONOMY OF LAW 287, 289 (Robert George ed., 1996) [hereinafter Coleman, *Authority and Reason*] (defining "incorporationism" as the claim that a rule of recognition can incorporate the community's morality into its law).

reasonable care in tort law, or unconscionability as a defense to the enforcement of contracts, have long since been incorporated into legal reasoning; they have social pedigrees.[85] A familiar example of inclusive positivist analysis would be the conventional practice of seeking to determine whether, for the purposes of the Eighth Amendment, it is *really* cruel and unusual to execute children or people with mental disabilities. Supreme Court decisions that refer to evolving standards of decency and consider the proportionality of capital punishment and the culpability of the offender are examples of inclusive positivist reasoning.[86] In jurisprudential terms, the moral evaluation that, say, executing children is cruel and unusual is a proposition of law as well as morality *in virtue of* a conventional practice of incorporating that evaluation into legal judgments.

Exclusive positivists, by contrast, maintain that legal validity "cannot turn on matters of substance, moral or other."[87] What about the judge deciding an Eighth Amendment case, and concluding that executing juveniles is cruel and unusual? An exclusive positivist judge may also look to moral standards, such as decency and culpability, when deciding an Eighth Amendment issue. The difference is that the exclusive positivist judge would frankly concede that she is engaging in moral reasoning unless the principles upon which the decision rests have been "accepted and practiced by officials from the internal point of view."[88] In tort cases, for example, there is a conventional rule permitting judges to make reference to moralized conceptions of reasonableness. An exclusive positivist would have no difficulty concluding that, for example, it was unreasonable for a foreign-exchange program not to monitor more closely a high-school student's relationship with her host family to ensure that no sexual misconduct was occurring in the home.[89] The only real difference between inclusive and exclusive positivism is that the inclusive positivist judge would contend that *the law* permits reference to reasonableness in the moral sense, while an exclusive positivist would claim that her role as judge permits her sometimes to do what is morally the right thing, even if the law does not prescribe one resolution or the other.

Different flavors of anti-positivism of course involve more direct connections between law and morality. An anti-positivist would contend

[85] *See* HART, *supra* note 31, at 271.

[86] *See, e.g.,* Roper v. Simmons, 543 U.S. 551 (2005); Atkins v. Virginia, 536 U.S. 304 (2002).

[87] Coleman, *Authority and Reason*, *supra* note 84, at 290.

[88] SHAPIRO, *supra* note 27, at 269.

[89] *See* Beul v. ASSE Int'l, Inc., 233 F.3d 441 (7th Cir. 2000).

that, in order to ascertain the content of the law, it is necessary to "point to some normative facts alongside the social facts."[90] Classical natural law theory directs a judge to consider whether a would-be law is in fact an "ordinance of reason for the common good, made by [a ruler] who has care of the community."[91] Dworkin's distinctive anti-positivist position presents judges with the herculean task of determining how a decision should fit with past political acts, including the enactment of statutes by the legislature and prior judicial decisions, while also explaining how the decision "figure[s] in or follow[s] from the principles of justice, fairness, and procedural due process that provide the best constructive interpretation of the community's legal practice."[92] In his now somewhat shopworn example of *Riggs v. Palmer*, the beneficiary of a will was denied the right to obtain his bequest because he had murdered the testator.[93] The court in that case based its decision on the principle, "no person should profit from his or her own wrongdoing," which is a reasonable enough maxim of morality, but cannot be subsumed under a Hartian rule of recognition.[94] Dworkin argues that "legal principles exist which determine the right answer to the legal question at issue."[95] It is essential to understand that he believes *Riggs*, the murdering-heir case, to have a right answer *as a matter of law*. Judges do not have the discretion to decide the case according to their own beliefs about morality, because they do not look only to social facts to determine the right resolution of disputes, but to the morality of the political community of which they are a part. Because judges rely on non-legal standards in deciding cases, positivism cannot be an apt description of law.[96] Evaluative judgments are required in the dimension of fit as well as, more obviously, in the dimension of justification, and these judgments

[90] Hershovitz, *supra* note 84, at 1166.

[91] THOMAS AQUINAS, AQUINAS ON LAW, MORALITY, AND POLITICS 10 (William P. Baumgarth et al. eds., Richard J. Regan trans., 2d ed. 2002).

[92] RONALD DWORKIN, LAW'S EMPIRE 225 (1986).

[93] Riggs v. Palmer, 22 N.E. 188, 188–89 (N.Y. 1889).

[94] Scott J. Shapiro, *On Hart's Way Out*, *in* READINGS IN THE PHILOSOPHY OF LAW 125, 152 (Jules Coleman ed., 2013).

[95] SHAPIRO, *supra* note 27, at 263.

[96] According to Dworkin, positivism is necessarily *social-facts positivism*, which is committed to an austere metaphysics in which all law is necessarily determined by social facts alone. *See id.* at 266. Varieties of positivism that are not committed to a model of social facts may be able to avoid this critique. *See, e.g.*, Benjamin C. Zipursky, *The Model of Social Facts*, *in* HART'S POSTSCRIPT: ESSAYS ON THE POSTSCRIPT TO THE *CONCEPT OF LAW* 219 (Jules Coleman ed., 2001).

cannot be accounted for using the Hartian idea of conventionally practiced social rules.

Even if one believed that exclusive positivism offered the best theoretical account of the concept of law (e.g. for reasons given by Raz[97]), and also insisted, implausibly, that judges may not make reference to extra-legal moral considerations, AI systems must nevertheless contend with the inevitable presence of judgment in legal decision-making. Hart recognized that positivism was vulnerable to the superficial but nevertheless plausible charge that it entailed a formalistic or mechanical approach to adjudication. He famously distinguished between cases lying near the core of the settled meaning of a rule and those farther out within the rule's penumbra.[98] More generally, as a matter of logic and language, rules cannot provide for their own interpretation.[99] The application of legal rules depends on criteria of relevant that "depend on many complex factors," including the purpose that can be attributed to the rule;[100] the notion that when fashioning a rule it is impossible for a judge or legislator to foresee all the circumstances of its application in the future;[101] and the fact that any rule can be read broadly or narrowly, depending on standards such as the materiality of facts to the prior decision, which themselves are indeterminate.[102] Judges and lawyers have discretion in the application of rules, and the vice of formalistic legal theories is the attempt to disavow or disguise the necessity of the exercise of discretion.[103] None of these observations depends on the relevance of *moral* judgments to the content of law. But legal interpretation and application does, in all but the core of settled meaning of rules, require the exercise of judgment.

One of Dworkin's examples in *Law's Empire* illustrates the centrality of judgment, though not necessarily moral judgment, in legal decision-

[97] RAZ, *Authority, Law, and Morality*, *supra* note 28, at 210.

[98] HART, *supra* note 31, at 126.

[99] *Id.* ("Particular fact-situations do not await us already marked off from each other, and labelled as instances of the general rule . . . .").

[100] *Id.* at 127.

[101] *Id.* at 133.

[102] *Id.* at 130, 134.

[103] *Id.* at 129. Interestingly, given what was said above about the necessity of discretion along the axes of fit and justification, Shapiro sees Dworkin's critique of Hart, with its insistence that there are right answers to questions of law as an attempt "to salvage a rump version of formalism as a serious jurisprudential account." SHAPIRO, *supra* note 27, at 261. Dworkin contends that judges do not have discretion in the strong sense, because principles of the community's political morality determine the right answer to the legal question at issue. *Id.* at 263.

making. He writes about a decision from the English House of Lords permitting recovery for negligent infliction of emotional distress to a plaintiff who did not contemporaneously observe an accident causing serious physical harm to members of her family, but heard the news, and rushed to the hospital to see her injured loved ones.[104] American lawyers know of similar cases from the California Supreme Court, such as *Dillon v. Legg*[105] and *Thing v. LaChusa*.[106] In all of these cases, the result depends on the resolution of a number of considerations such as black-letter legal doctrine, such as the foreseeability of an injury; norms governing the interpretation of legal principles, such as the preference for bright line rules and *ex ante* certainty; and "policy" arguments such as the concern that emotional distress is easily feigned or exaggerated. Dworkin insists that a judge can decide that case only by finding the interpretation of precedent cases that yields a coherent set of principles about justice, fairness, and due process that are contained within the precedents,[107] and which shows the political history of the community in its best light, morally speaking.[108] One need not agree with Dworkin that the decision turns on matters of political morality to acknowledge that there is unlikely to be any higher-order principle that synthesizes the considerations bearing on the decision in a coherent way.[109] Some decisionmakers are likely to prefer the result that awards full compensation to the plaintiff for her emotional distress, either because of a judgment that the injury deserves compensation, or for the instrumental reason that allowing recovery for emotional distress damages will further the deterrent function of tort law. Other judges may give greater weight to the rule-of-law considerations of limiting the discretion of judges and juries by insisting on a clear, easily administrable rule (such as the requirement that the plaintiff have been physically present at the scene of

[104] *See* DWORKIN, *supra* note 83, at 23-29 (discussing McLoughlin v. O'Brian [1983] 1 AC 410 (Eng.)).

[105] 441 P.2d 912 (Cal. 1968).

[106] 771 P.2d 814 (Cal. 1989).

[107] DWORKIN, *supra* note 83, at 243.

[108] *Id.* at 248-49.

[109] Dworkin's analogy of a chain novel, in which successive judges are like the authors of successive chapters in a novel, may appear to make this type of decision more tractable for an AI system. As Stanley Fish has argued, however, preceding chapters in a chain novel cannot fully constrain what the novel will eventually become; subsequent interpreters always have freedom to, for example, characterize what has gone before as a social satire or a comedy of manners. Stanley Fish, *Working on the Chain Gang: Interpretation in Law and Literature*, 60 TEX. L. REV. 551, 554 (1982). The existence of precedent, in law or literature, does not do away with the role of judgment.

the accident and closely related to the victim), as opposed to a more open-ended standard such as whether severe emotional distress was reasonably foreseeable.

At least as the technology currently exists, AI systems do not cope well with multiple sources of legal authority (including materials such as the interpretive rules, guidance, and policy statements that typify administrative governance), persuasive authority such dicta and out-of-state law, analogies between cases that are not factually similar but which would be recognized by a human decisionmaker, the wide range of factual variations among legally similar cases, and anomalous cases that do not fit the patterns on which the system was trained.[110] An AI system would have considerable difficulty replicating human decision-making in a case like Dworkin's example, which is something any first-year student in torts should be able to deal with easily. However, the point of this example is not to critique the state of existing technology; rather, it is to show that hard cases are not hard for computers because of any specifically *moral* content to law. As a matter of jurisprudence, it tells against Dworkin's example that an appellate judge deciding the case would reason in exactly the same manner, regardless of her prior commitment to exclusive positivism, inclusive positivism, or Dworkin's theory of law as integrity. Thus, any gap between human and robot lawyers and judges revealed by cases like this does not turn on the relationship between law and morality.

B. Legal Prediction vs. Legal Authority

The obvious response to this pattern of argument is to differentiate the resolution of hard or marginal cases from the great majority of legal decisions made by lawyers and judges. Even Hart, for whom a conventional rule of recognition is central, concedes that sometimes the law may run out, in which case judges are forced to legislate in the gaps, creating new law.[111] After a case like the one described above is resolved by the state's highest court, a few application questions may remain, but eventually cases decided by lower courts fall into relatively predictable patterns. Certain categories of plaintiffs will be deemed sufficiently closely related to the victim to be entitled to recover for emotional distress, and it will become clear what counts as "close enough" to the accident scene to make the distress reasonably foreseeable. An AI system can be trained on a data set consisting of decided lower-court cases, and after the law has had an

[110] Pasquale & Cashwell, *supra* note 18, at 42-44.

[111] *See* HART, *supra* note 31, at 145.

opportunity to develop, will probably do a pretty good job predicting the outcomes of cases given facts about the relationship between the plaintiff and the victim, and the location of the plaintiff relative to the accident site.

But notice a tacit premise in this response, which is pervasive when thinking about AI systems potentially replacing lawyers. The premise is that a legal judgment, which would serve as the underpinning of legal advice to a client concerning the lawfulness of a proposed course of action, is nothing more than a *prediction* concerning how courts would resolve the issue if it were litigated. This premise has a long history within American jurisprudence. Early in his career, Karl Llewellyn wrote that "[w]hat . . . officials do about disputes is, to my mind, the law itself."[112] And Oliver Wendell Holmes, Jr., notoriously told an audience of law students that if their clients asked what the law required them to do, they were *really* asking about the likelihood of detection and punishment.[113] If the law really is nothing more than a prediction, then we have probably already arrived at a time when technology is capable of equaling or surpassing the performance of human lawyers at predicting legal judgments. Legal decision-support systems already offer fast, accurate analysis of the expected results of particular types of motions before particular judges. But those predictions do not, and cannot, have the status of law. They are suggestions of what the law might be, but the law, by its nature, is a means for humans to offer reasons to one another, in response to the circumstances of encountering one another as equal and mutually accountable.[114] The law is a means for giving the types of reasons that human moral agents owe to one another, in response to others' demands for accountability.

I have given a more elaborate defense of this view elsewhere.[115] Briefly summarizing for present purposes, the argument is that morality is a matter of "what we owe each other."[116] What we owe to each other is accountability. We have the standing to address demands to one another, to

[112] KARL N. LLEWELLYN, THE BRAMBLE BUSH: THE CLASSIC LECTURES ON THE LAW AND LAW SCHOOL 5 (1930).

[113] *See* O. W. Holmes, *The Path of the Law*, 10 HARV. L. REV. 457, 459-62 (1897) (setting out the so-called "bad man" theory of law). *But see* David Luban, *The Bad Man and the Good Lawyer: A Centennial Essay on Holmes's* The Path of the Law, 72 N.Y.U. L. REV. 1547 (1997) (arguing that Holmes's lecture is frequently misunderstood).

[114] DARWALL, *supra* note 32, at 101.

[115] *See, e.g.*, Wendel, *supra* note 41; W. Bradley Wendel, *Fiduciary Theory and the Capacities of Clients: The Problem of the Faithless Principal*, PENN. ST. L. REV. (forthcoming 2019); W. Bradley Wendel, *The Limits of Positivist Legal Ethics: A Brief History, a Critique, and a Return to Foundations*, 30 CAN. J.L. & JURIS. 443 (2017).

[116] *See generally* SCANLON, *supra* note 34.

either act or refrain from acting in some manner that affects our interests, or to give reasons for refusing that demand. In Stephen Darwall's example of an everyday relationship of authority, one person says to another, "hey, you're stepping on my toe and it hurts—move your foot!"[117] This seemingly simple example actually reveals something deep and important about the grounds of moral obligations. The fundamentally intersubjective nature of demand for accountability, issued by one free and equal person to another, places conditions on the types of reasons that may be given in response. Those reasons are not external—that is, they do not pertain to outcomes or the state of the world; rather, they must be *relational*.[118] Why? Because the authority to demand accountability presupposes that addresser and addressee of reasons share a point of view.[119] As Christine Korsgaard has argued, the authority of morality proceeds from our practical identity that gives rise to reasons and obligations.[120] Valuing ourselves under a description—a practical identity—involves recognizing the reasons that we share with others. As Korsgaard puts it, the reasons of others have a standing with ourselves on a par with the reasons we have.[121] Addressing a demand for accountability to another forces that person to acknowledge the value of the addresser's humanity, and thus her status a self-originating source of value.[122]

The further premise, connecting law with the relationship of authority and accountability among free and equal persons, is that law provides a means for making and complying with these demands for accountability in a complex, pluralistic society. People who seek to cooperate and engage in mutually beneficial activities require some way of not only coordinating action but also acknowledging others' entitlement to be treated as free and equal—that is, their second-personal authority. Law has the moral aim of rectifying the problems of a community in which people demand accountability from others but are prevented by uncertainty and disagreement, or even simply the scale and complexity of a modern society, from giving sufficient reasons in the kind of idealized face-to-face

[117] DARWALL, *supra* note 32, at 5-7.

[118] *Id.* at 246-47. This observation is related to, and helps support, the arguments given by Russ Pearce and Eli Wald for a relational conception of the lawyer-client relationship. *See* Eli Wald & Russell G. Pearce, *Being Good Lawyers: A Relational Approach to Law Practice*, 29 GEO. J. LEGAL ETHICS 601, 616-18 (2016).

[119] *See* DARWALL, *supra* note 32, at 249.

[120] CHRISTINE M. KORSGAARD ET AL., THE SOURCES OF NORMATIVITY 101 (Onora O'Neill ed., 1996).

[121] *Id.* at 140.

[122] *Id.* at 143.

encounter imagined by moral philosophers.[123] The authority of law, which is to say its capacity to furnish reasons that satisfy the moral demand for accountability, depends on its serving the human need to comply with the requirements of morality.[124]

AI systems may assist lawyers in carrying out their obligations to clients and non-clients. It would undoubtedly be useful in many cases for a lawyer to have a clear picture of how different judges decide motions on various grounds. At the heart of the social role of lawyer, however, is the function of supporting the moral aim of the law, which is to furnish reasons that satisfy the demand for accountability that exists in a society of free and equal persons. What I have referred to as the core lawyering function is involved when lawyers (1) advise clients that the law permits them to undertake some course of action (the compliance role); (2) draft contracts, deeds, wills, trust instruments, or other documents that have the effect of altering the legal rights of clients and those with whom they interact (the private ordering role); counseling clients on how best to achieve their objectives within the limitations of the law (the planning role); and (5) certify expressly or implicitly to a court that they are asserting a position in litigation that has adequate factual and legal support (the role as officer of the court). Predictions of future judicial decisions and summaries of precedents may assist lawyers in carrying out these tasks. But the distinctive contribution of lawyers to the process is the judgment that a client has a legally sufficient reason for an action. Lawyers manifest the authority of law in their interactions with clients and others by giving reasons that addressees can acknowledge from the standpoint shared by free and equal persons.[125] Here is a crucial point that sometimes goes unrecognized in theoretical legal ethics: The practical authority of law depends on its being administered by public and quasi-public officials—that is, judges and lawyers—who are subject to requirements of professional role morality that orient their actions to the moral aim of law.

[123] SHAPIRO, *supra* note 27, at 170-72 (defending moral aim thesis).

[124] *See* RAZ, MORALITY OF FREEDOM, *supra* note 29, at 56-59 (defending service conception of authority).

[125] *See* Waldron, *supra* note 33, at 26-27 (legally sufficient reasons are those which address "people's capacities for practical understanding, for self-control, and for the self-monitoring and modulation of their own behavior, in relation to norms that they can grasp and understand").

Law is not self-administering or self-interpreting. It depends for its efficacy and therefore its authority on appropriate conduct by human agents.[126]

This observation suggests the response to an objection that may have occurred to many readers. The functional account of law set out here relies on an encounter between two people who meet, literally or metaphorically in the public square, and establish a relationship of authority and accountability. A political community can be seen as a massively scaled-up version of Darwall's example of people stepping on each other's toes. Those people require some means of making authoritative demands and responding with the right sorts of reasons. However, it is not obvious that the law is the only means available to people who encounter each other and need to deal with disagreements. I have argued that law is a technology, or a means for addressing and responding to demands for accountability. But other technologies are available. What difference does it make if two people decide to resolve a dispute by flipping a coin, consulting chicken entrails, or using a Magic 8-Ball?[127] And, if they can use a Magic 8-Ball, what would be wrong with using some kind of AI-enabled online dispute resolution system? There is no need for that system to have all of the properties and virtues of human law if it is merely a more elaborate version of any other system two people might employ to resolve disputes.

The response to this objection is that most of law does not work like a case in which two parties consent to resolve a dispute using a coin toss, the Magic 8-Ball, or an online dispute resolution system. The relationship between consent and authority is a bit subtle, and confusion may result from the influence of the conception of authority defended by Joseph Raz. He gives the example of "two people who refer a dispute to an arbitrator."[128] The arbitrator resolves the dispute and gives reasons supporting the decision. Importantly, the arbitrator's decision should be based on the reasons that apply in any event to the disputing parties. That is not to say that the parties may challenge the arbitrator's decision as not

[126] *See* RAZ, AUTHORITY OF LAW, *supra* note 27, at 9 (noting that authority must be both effective and under a claim of right to be distinguished from brute force).

[127] For readers not of a certain age, the Magic 8-Ball is a toy that purports to tell fortunes. The user shakes up a plastic ball full of liquid and looks through a window at a twenty-sided plastic piece floating inside. "Fortunes" are printed on each face of the twenty-sided piece. Thus, you ask the 8-Ball a question and it will respond with guidance such as "It is certain," "Outlook good," or "My sources say no." *See Magic 8-Ball*, WIKIPEDIA (Feb. 18, 2019, 20:17 UTC), https://web.archive.org/web/20190315163848/https://en.wikipedia.org/wiki/Magic_8-Ball.

[128] RAZ, MORALITY OF FREEDOM, *supra* note 29, at 41-43.

correctly reflecting the balance of preexisting reasons.[129] In fact, the whole point of seeking an authoritative decision is that the arbitrator's decision furnishes a new reason that replaces or preempts the reasons brought by the parties to the dispute. The authoritative decision of the arbitrator is dependent upon the parties' preexisting reasons, but preempts those dependent reasons. "[R]easons that could have been relied upon to justify action before [the arbitrator's] decision cannot be relied upon once the decision is given."[130] This dual relationship of dependence and preemption is central to the Razian picture of authority.

I very much intend to rely on Raz's account of legal authority, as I have done in previous work.[131] It is the linchpin of the conceptual argument against the possibility of robo-lawyers or robo-judges fully displacing human agents in those roles. Nevertheless, it is important to acknowledge a potentially misleading feature of Raz's arbitration example. The authority of an arbitrator requires the consent of the parties. In the U.S., the Federal Arbitration Act makes arbitration fundamentally a matter of contract; the Supreme Court has been very willing to enforce arbitration provisions in contracts, even if the practical consequence is to preclude the litigation of certain types of claims.[132] Much law, by contrast, operates non-consensually. Practical authority, by definition, means altering the normative situation of its subjects. It is easy to understand an arbitrator's authority given consensually as the autonomous choice of a rational agent to agree to abide by the decision of a neutral party.[133] Law, however, often operates non-consensually, even coercively.[134] The social contract tradition, associated with thinkers like John Locke, and made familiar in American political ideals through the language of the Declaration of Independence, holds that political authority depends on the consent of the governed.[135] As Hume and countless other critics have pointed out, however, no government of a modern state has been founded without some "usurpation or

[129] *Id.* at 47.

[130] *Id.* at 42

[131] *See* W. BRADLEY WENDEL, LAWYERS AND FIDELITY TO LAW 108-12 (2010).

[132] *See, e.g.*, Am. Express Co. v. Italian Colors Rest., 570 U.S. 228, 233-35 (2013) (upholding waivers of class action rights as part of consent to arbitration in merchant-card issuer agreement).

[133] *See* BEAUCHAMP & CHILDRESS, *supra* note 46, at 122-23 (explaining the normative force of informed consent as the individual's autonomous authorization).

[134] *See generally* Robert M. Cover, *Violence and the Word*, 95 YALE L.J. 1601 (1986) (providing the classic account of the inherent violence of the law).

[135] *See* JOHN LOCKE, SECOND TREATISE OF CIVIL GOVERNMENT, ch. IV, § 22 (J.W. Gough ed., Basil Blackwell 1948) (1690).

conquest."[136] With the exception of naturalized citizens who take an oath of allegiance to their adopted home, most residents of a nation-state never give express consent to be bound by the state's laws; they are simply born within the state's territorial jurisdiction. Locke responded by attempting to infer *tacit* consent from a wide variety of ways in which people actively or passively participate in a society.[137] However, tacit consent theories have not fared much better in the history of debates over political authority. It is generally accepted that mere acquiescence to authority does not constitute consent unless the person purportedly giving consent has a readily available means of dissent which can be performed at little cost, will be respected by the would-be authority, and will not be an occasion for retaliation against the dissenter.[138] People may obey law out of habit or a sense of reverence that has been inculcated through mandatory rituals of conformity, such as reciting the Pledge of Allegiance in school as children. Some may be attracted to the slogan "America—love it or leave it," but most American citizens do not have a meaningful exit option due to a lack of financial resources, language proficiency, or job skills that would permit them to emigrate from their native country (to say nothing of the substantial burden of leaving behind one's family and community to live as an expatriate).

Raz himself is a skeptic regarding legal authority.[139] The argument briefly summarized above goes beyond what Raz himself would accept, but I believe it is a sound case for the authority of law in a political community that perceives the need to establish means to plan and coordinate conduct when individuals are unable to resolve conflict due to complexity, uncertainty, and pluralism.[140] The important point is that law, by its nature, claims authority. If one believes that law cannot make good on this claim, then there is little to distinguish a human lawyer from any other human who makes predictions about when coercive means will be applied to force someone to do something. As Hart showed in his critique of Austin,

[136] David Hume, *Of the Original Contract*, *reprinted in* SOCIAL CONTRACT 147, 151 (Ernest Barker ed., 1947).

[137] LOCKE, *supra* note 135, at ch. VIII, § 119.

[138] *See* A. JOHN SIMMONS, MORAL PRINCIPLES AND POLITICAL OBLIGATIONS 80-81 (1979).

[139] *See* RAZ, AUTHORITY OF LAW, *supra* note 27, at 233 (arguing that "there is no obligation to obey the law even in a good society whose legal system is just").

[140] *See* SHAPIRO, *supra* note 27, at 172 ("Given the complexity, contentiousness, and arbitrariness of modern life, the moral need for plans to guide, coordinate, and monitor conduct are enormous. Yet, for the same reasons, it is extremely costly and risky for people to solve their social problems by themselves, via improvisation, spontaneous ordering, or private agreements, or communally, via consensus or personalized forms of hierarchy.").

however, there can be no legal obligation without the acceptance of the law as creating obligations, and therefore reasons—that is, serving as a practical authority.[141] If one sees the function of law as enabling individuals to give reasons to others, in response to a demand for accountability, then the role of lawyers becomes clear: They serve clients by assisting them in providing reasons that appropriately refer to the public authorization of their actions by law.[142]

IV. Conclusion

At least as the technology currently exists, no artificial system has the standing that is presupposed by the giving of legal reasons. AI may someday be able to emulate or model human moral reasoning, but what it can never do is *be* a free and equal person, in a second-person relationship with another free and equal person. Without the relationship of accountability and authority, the law cannot create obligations and reasons for action. The capacity of legal rules and principles to furnish reasons, create obligations, and possess authority all depends on the shared standpoint of mutual respect adopted by free and equal persons. A computer system or a robot lawyer could perhaps emulate a human lawyer, but it would lack the authority necessary to function as a lawyer. The core lawyering function, by its nature, is suitable only for human agents. Lawyers do many things outside their core function, however, and the legal profession has already witnessed disruption from the replacement of human lawyers with artificial systems in tasks like privilege reviews in discovery. The claim in this paper is unlikely to provide comfort to lawyers facing displacement in these non-core functions. But a world of robot lawyers is not on the horizon either—not because of limitations of existing or foreseeable technology, but owing to the dependency of law on relationships of authority and accountability.

[141] *See* HART, *supra* note 31, at 88-91 (discussing how rules from internal point of view create obligations, while from external point of view they seem only to provide possibility of punishment); *see also* GERALD J. POSTEMA, LEGAL PHILOSOPHY IN THE TWENTIETH CENTURY: THE COMMON LAW WORLD 291-99 (2011) (volume 11 of a series titled *A Treatise of Legal Philosophy and General Jurisprudence*) (explaining Hart's concept of the internal point of view); Scott J. Shapiro, *What Is the Internal Point of View?*, 75 FORDHAM L. REV. 1157, 1157 (2006) ("The internal point of view is the practical attitude of rule acceptance—it does not imply that people who accept the rules accept their moral legitimacy, only that they are disposed to guide and evaluate conduct in accordance with the rules.").

[142] *See generally* Wendel, *supra* note 41.

It may be that technology will advance one day to the point that artificial general intelligence exists that is reflectively self-conscious, values itself as a rational being, understands that other rational beings have value in the same way, and shares a practical evaluative standpoint which can serve as the foundation of reasons given to other rational beings in justification of actions that affect their interests. The theoretical foundations of the position defended here are basically Kantian. What matters for Kant is pure rational agency, and that may be something that computers eventually achieve. If that happens—if AI systems become self-originating sources of value and bearers of dignity—then there would no longer be any objection to them serving as lawyers or judges.