DePaul Law Review

Volume 73 | Issue 2 Article 7

AI Malpractice

Bryan H. Choi

AI MALPRACTICE

*Bryan H. Choi**

Should AI modelers be held to a professional standard of care? Recent scholarship has argued that those who build AI systems owe special duties to the public to promote values such as safety, fairness, transparency, and accountability. Yet, there is little agreement as to what the content of those duties should be. Nor is there a framework for how conflicting views should be resolved as a matter of law.

This Article builds on prior work applying professional malpractice law to conventional software development work, and extends it to AI work. The malpractice doctrine establishes an alternate standard of care—the customary care standard—that substitutes for the ordinary reasonable care standard. That substitution is needed in areas like medicine or law where the service is essential, the risk of harm is severe, and a uniform duty of care cannot be defined. The customary care standard offers a more flexible approach that tolerates a range of professional practices above a minimum expectation of competence. This approach is especially apt for occupations like software development where the science of the field is hotly contested or is rapidly evolving.

Although it is tempting to treat AI liability as a simple extension of software liability, there are key differences. First, AI work has not yet become essential to the social fabric the way software services have. The risk of underproviding AI services is less troublesome than it is for conventional professional services. Second, modern deep-learning AI techniques differ significantly from conventional software development practices, in ways that will likely facilitate greater convergence and uniformity in expert knowledge.

Those distinguishing features suggest that the law of AI liability will chart a different path than the law of software liability. For the immediate term, the interloper status of AI indicates a strict liability approach is

* Associate Professor of Law and Computer Science & Engineering, The Ohio State University. I thank Ruth Colker, Olwyn Conway, Rebecca Crootof, Bridget Dooling, John Goldberg, Margaret Kwoka, Stephan Landsman, Grace Li, Catherine Sharkey, Ric Simmons, Rebecca Wexler, and Patti Zettler, and attendees of the Clifford Symposium on Tort Law & Social Policy, and the BU Cyber Alliance Speaker Series for helpful input at early stages. I also thank Michael Adamo, Kevin Gibbons, Joseph van't Hooft, Joshua Menden, Damini Mohan, Rama Naboulsi, Courtney Pyatt, and Dane Sowers for excellent research assistance. This work was supported in part by NSF CCF-2131531.

most appropriate, given the other factors. In the longer term, as AI work becomes integrated into ordinary society, courts should expect to transition away from strict liability. For aspects that elude expert consensus and require exercise of discretionary judgment, courts should favor the professional malpractice standard. However, if there are broad swaths of AI work where experts can come to agreement on baseline standards, then courts can revert to the default of ordinary reasonable care.

Table of Contents

Introduction

As the tide of AI discourse shifts from AI performance to AI safety,[1] and from ethics to law, most legal efforts have focused on *ex ante* risk-management approaches. These approaches have been criticized as being underdeveloped, ill-suited to the problem, and lacking in *ex post* measures offering civil recourse.[2] A greater role for judicial common

1. *See* Press Release, U.S. Dep't of Com., At the Direction of President Biden, Department of Commerce to Establish U.S. Artificial Intelligence Safety Institute to Lead Efforts on AI Safety (Nov. 1, 2023), https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial [https://perma.cc/43AE-329Y]; Noam Kolt, *Algorithmic Black Swans*, 101 Wash. U. L. Rev. (forthcoming 2024) (manuscript at 23) (describing the "growing interest and investment in AI safety").

2. *See* Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. Rev. 1347, 1378–79 (2023) ("This, then, is one of the core challenges for AI risk regulation: it deploys a largely ex ante regulatory tool best suited for readily quantifiable harms to address big, often-unquantifiable, often-contested, often-contextual, and often-individualized 'risks.' . . . Nor do these frameworks contemplate whether there exists an adequate backstop of tort liability through which individualized harms might get addressed and remedied."); Charlotte A. Tschider, *Medical Device Artificial Intelligence: The New Tort Frontier*, 46 BYU L. Rev. 1551, 1591–93, 1603 (2021) (arguing that ex post tort solutions are needed to regulate AI software-enabled medical devices); Bryan H. Choi, *NIST's Software Un-Standards*, Lawfare, at 30 (2024) ("[M]uch about [NIST's] AI Framework remains underdetermined. The AI Framework is voluntary. It is neither 'a checklist' nor 'an ordered

law is needed,[3] but there is considerable uncertainty about the path it should take.

Thus far, public oversight is being routed through private governance and self-regulation mechanisms. In the United States, Congress has proposed a smattering of legislative proposals that have failed to advance.[4] Instead, the leading effort has been the AI Risk Management Framework, issued by the National Institute of Standards and Technologies (NIST), which invites enterprises to engage in voluntary self-assessments of risk.[5] The European Union has enacted the EU AI Act, which categorizes AI systems as "limited-risk," "high-risk," or "unacceptable risk," and then seeks to calibrate compliance obligations accordingly.[6] Those obligations have been outsourced to private standard-setting organizations and have yet to be written.[7]

In turn, the AI community continues to view self-regulation through the lens of professional "ethics."[8] Major software companies, nonprofits, and academic institutions are engaged in efforts to develop ethical principles that address issues of fairness, accountability, transparency, manipulation, and more.[9] Universities are working on integrating ethics

---

set of steps.' Evaluations of its effectiveness are unknown and 'will be part of future NIST activities.' Moreover, its scope is extraordinarily broad, even relative to other NIST frameworks.").

3. *See* Mariano-Florentino Cuéllar, *A Common Law for the Age of Artificial Intelligence: Incremental Adjudication, Institutions, and Relational Non-Arbitrariness*, 119 Colum. L. Rev. 1773, 1779–80 (2019) (defending the role of common law as a "regulatory backstop" or "first-draft regulatory framework," and "the centrality of reasoned deliberation across people and institutions" in the American tradition).

4. *See* Kolt, *supra* note 1 (manuscript at 36–38) (describing and critiquing recent Congressional bills that would delegate responsibility to the Federal Trade Commission (FTC) or a newly created National AI Commission).

5. *See* Choi, *supra* note 2, at 29.

6. *See* Clara Hainsdorf et al., *Dawn of the EU's AI Act: Political Agreement Reached on World's First Comprehensive Horizontal AI Regulation*, White & Case (Dec. 14, 2023), https://www.whitecase.com/insight-alert/dawn-eus-ai-act-political-agreement-reached-worlds-first-comprehensive-horizontal-ai [https://perma.cc/HRA6-6VZC].

7. *See* Kaminski, *supra* note 2, at 1402 (noting that the EU AI Act "places big, complex, and contested policy decisions in the hands of private entities"); Kolt, *supra* note 1 (manuscript at 30).

8. *See* Partnership on AI, 2019 Annual Report 2 (2020), https://www.partnershiponai.org/wp-content/uploads/2021/01/PAI-2019-Annual-Report.pdf [https://perma.cc/3PB2-PLZH] (reporting that membership grew to over one hundred partner organizations spanning thirteen countries and four continents).

9. *Id.* at 6 (listing key focus areas to include fairness, transparency, and accountability in algorithmic decision-making, as well as AI-generated mis/disinformation in the media ecosystem); Scope, FAT/ML 2014 Conference, https://www.fatml.org/schedule/2014/page/scope-2014 [https://perma.cc/HNK4-Q6RJ] ("This interdisciplinary workshop will consider issues of fairness, accountability, and transparency in machine learning. It will address growing anxieties about the role that machine learning plays in consequential decision-making in such areas as commerce, employment, healthcare, education, and policing."); Virginia Dignum et al., *Ethics by Design: Necessity or Curse?*, 2018 Proc. AAAI/ACM Conf. on AI Ethics & Soc'y 60, 61–62 (summarizing conference proceedings describing accountability, responsibility, and transparency as the three principles

more tightly within the computer science curriculum.[10] While those efforts have generated attention and momentum,[11] critics have pointed out that "[e]thics as a construct is notoriously malleable and contested" and that "ethics lacks a hard enforcement mechanism."[12] More cynical voices have warned that the focus on ethics is a deliberate ploy by interested parties to avoid true oversight.[13]

Law would provide a harder enforcement mechanism, if lawmakers could formulate appropriate liability rules for AI harms. Yet, with modern AI techniques such as deep neural networks, there is broad consensus that failures are difficult to diagnose or explain in human-understandable terms. Thus it can be difficult to discern when law should exact remedies from AI modelers versus when law should let costs lie where they fall. Moreover, many policymakers remain reluctant to err on the side of overdeterrence, because the perceived potential of AI technologies to improve public safety, health, and welfare is so seismic.[14]

Conventional approaches to AI liability attempt to simplify the problem by treating AI systems as self-contained black boxes. For example, one popular branch of commentary focuses on the "intelligence" part of

---

that are key to "guaranteeing ethical behavior 'by design,' i.e., embedded in the [AI] system's implementation").

10. *See* Christina Pazzanese, *Trailblazing Initiative Marries Ethics, Tech*, Harv. Gazette, (Oct. 16, 2020) (describing a recent curricular initiative at Harvard called Embedded EthiCS, "a groundbreaking novel program that marries the disciplines of computer science and philosophy," motivated by the premise that "the surest way to get the industry to act more responsibly is to prepare the next generation of tech leaders and workers to think more ethically about the work they'll be doing").

11. *See* White House Office of Sci. & Tech. Policy, Blueprint for an AI Bill of Rights (Oct. 2022); Press Release, U.S. Dep't of Def., DOD Adopts Ethical Principles for Artificial Intelligence (Feb. 24, 2020), https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/ [https://perma.cc/BW72-6WZU].

12. Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. Davis L. Rev. 399, 408 (2017); *see also* Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 Nature Mach. Intel. 501 (2019); Sanna J. Ali et al., *Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk among Ethics Entrepreneurs*, 2023 ACM Conf. on Fairness, Accountability & Transparency 217, 220 (finding that ethics workers face numerous challenges including a lack of "buy-in" from leadership).

13. *See* Rebecca MacKinnon, *"Ethics" and "AI": Can We Use These Terms to Take Effective Action?*, Blog L.A. Rev. Books (Feb. 22, 2020), https://blog.lareviewofbooks.org/provoations/ethics-ai-can-use-terms-take-effective-action [https://perma.cc/U9A2-LFQ9] ("[A]cademics have cautioned against 'ethics-shopping,' 'ethics-washing,' and the development of vague ethical principles as an effort by companies to escape regulation and accountability. Others have accused industry of manipulating academia by funding 'ethical AI' research programs that focus on self-regulatory standards in order to avoid 'legally enforceable restrictions of controversial technologies.'").

14. *See* Off. of Sci. & Tech. Pol'y, White House, American Artificial Intelligence Initiative: Year One Annual Report 13 (2020), https://www.whitehouse.gov/wp-content/uploads/2020/02/American-AI-Initiative-One-Year-Annual-Report.pdf ("Not using AI technologies because of perceived or potential harms, however, could be just as problematic, depriving individuals—or the Nation—of the significant benefits that AI technologies could bring.").

"artificial intelligence" to argue that sufficiently strong AI systems could be given status equivalent to human or other sentient actors.[15] As long as AI behavior conforms to or exceeds that of the ordinary reasonable person, perhaps no liability should attach.[16] Another set of commentary focuses instead on the "artificial" aspects of AI, and analyzes such systems like ordinary manufactured products.[17] Under this approach, liability would be assessed according to whether the AI system has a "defect" that is unreasonably dangerous. Both approaches take the tack of evaluating the AI system as a discrete entity or object, while minimizing the human processes behind the development of that AI system. While that mental shortcut works best when the causal mechanisms are intuitive and simple to understand, it is less helpful when they are counterintuitive or complex.

As an alternative approach, David Lehr and Paul Ohm have called for lawmakers to open the black box and investigate the actual work involved in making AI systems.[18] Doing so would "advance contemporary debates about machine learning" by dispelling the assumption

---

15. *See* Ryan Abbott, The Reasonable Robot: Artificial Intelligence and the Law 9 (2020) (proposing that tort law should "treat AI like a person and focus on the AI's act rather than its design"); David Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 Wash. L. Rev. 117, 145, 150 (2014) ("A machine that can define its own path, make its own decisions, and set its own priorities may become something other than an agent. . . . Conferring 'personhood' on these machines would resolve the agency question . . . ."); *see also* Ryan Calo, *Robots as Legal Metaphors*, 30 Harv. J.L. & Tech. 209, 231 (2016); Patrick Hubbard, *"Sophisticated Robots": Balancing Liability, Regulation, and Innovation*, 66 Fla. L. Rev. 1803, 1862–65 (2014) (exploring doctrinal analogies of robots to employees, children, or animals). *But see* Harry Surden, *Artificial Intelligence and Law: An Overview*, 35 Ga. St. U. L. Rev. 1305, 1308–09 (2019) (pointing out that "Artificial General Intelligence" or "Strong AI" is unlikely to appear anytime soon).

16. *See* Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 Calif. L. Rev. 1611, 1653, 1679, 1686 (2017) (suggesting no liability if an autonomous vehicle meets a metric of being at least *twice* as safe as human performance).

17. *See* Bryant Walker Smith, *Automated Driving and Product Liability*, 2017 Mich. St. L. Rev. 1, 46 (stating that the central question in product liability claims against automated driving systems will be whether "either (a) a human driver or (b) a comparable automated driving system could have done better under the same circumstances"); *id.* at 44 n.213 (collecting sources examining product liability implications of automated driving); *cf.* Chinmayi Sharma & Benjamin C. Zipursky, *Who's Afraid of Products Liability? Cybersecurity and the Defect Model*, Lawfare (Oct. 19, 2023, 10:24 AM), https://www.lawfaremedia.org/article/who-s-afraid-of-products-liability-cybersecurity-and-the-defect-model [https://perma.cc/J4FJ-NWDB] (defending the law of products liability as having the flexibility and sophistication to address the challenges of software security lawsuits). *But see* Kenneth S. Abraham & Robert L. Rabin, *Automated Vehicles and Manufacturer Responsibility for Accidents: A New Legal Regime for a New Era*, 105 Va. L. Rev. 127, 142 (2019) (questioning whether the "defect" concept is worth retaining for fully automated vehicles, and proposing instead a no-fault regime).

18. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. Davis L. Rev. 653, 655 (2017) ("Our core claim is that almost all of the significant legal scholarship to date has focused on the implications of the running model . . . and has neglected most of the possibilities and pitfalls of playing with the data.").

that "black-box algorithms have black-box workflows."[19] In fact, they argue, the processes performed by AI modelers "are actually quite articulable."[20]

This Article takes up that invitation to dive into the details of how modern AI systems—and in particular deep neural networks[21]—are developed and deployed in the real world. By doing so, this Article seeks to shift the point of scrutiny from the AI system to the AI modeler, and to merge the active discourse on AI ethics with the burgeoning one on AI liability.

More specifically, this Article builds on prior work, *Software as a Profession*, which examined the interplay between tort liability and professional ethics.[22] There, the argument proceeded in three steps. First, I observed that professional ethics lack binding legal effect except for specially designated professions such as law or medicine. Second, I offered a new theoretical framework that explains why the law gives different treatment to such professionals, and how to decide which occupations are "professions" as a matter of law. In particular, I argue that the professional designation should be invested in occupations that: (1) necessarily employ subjective judgments because the field is not a precise science but an inexact art; (2) those subjective judgments carry high risk of bad outcomes; and (3) the occupation fulfills a vital societal function. Third, I showed why those factors apply to the work that software developers perform.

Here, I compare and contrast AI modeling to software development, and suggest that the differences are substantial enough to call into doubt whether AI modelers should be treated equivalently to software developers. While the professional malpractice framework may still prove appropriate for certain aspects of AI work, there is greater occasion to consider alternate regimes such as strict liability or ordinary negligence.

Part I offers a summary review of the professional malpractice doctrine. The remainder of the Article analyzes in turn each of the three "professional" factors as applied to AI. Part II offers a descriptive account of modern AI modeling work and the most salient ways in which it does or does not require the use of subjective judgments. In particular, most of the discretionary control occurs at the preparatory stages, including the choice of learning algorithm and initialization hyperparameters, plus certain aspects of data curation. Even here, AI modelers may not make

19. *Id.* at 657.
20. *Id.*
21. *See* Yann LeCun et al., *Deep Learning*, 521 Nature 436, 438 (2015).
22. *See* Bryan H. Choi, *Software as a Profession*, 33 Harv. J.L. & Tech. 557 (2020).

all those choices themselves, but instead reuse or rely on choices made by others. Part III restates the unavoidability of AI-based harm. Finally, Part IV reserves judgment on whether or when AI-based services will become an essential pillar of social fabric. Looking ahead, this Article seeks to tie the AI ethics literature closer to the law of professional malpractice, and predicts that the gravitas of the AI ethics movement will turn on the threshold legal question of whether AI work is deemed a profession, or merely a skilled occupation.

I. The Professional Malpractice Framework

The professional malpractice doctrine operates as an alternative to the ordinary negligence regime. Understanding why an alternative is needed at all requires a theory of what sets professions apart from ordinary occupations. In earlier work, I explained that this alternative is appropriate when there is simultaneously a need for legal oversight plus a need for deference to professional judgment over lay opinion.[23] In such circumstances, it is helpful to chart a middle path between ordinary negligence, safe harbor immunity, and enterprise liability.

The key, blackletter difference is the substitution of the customary care standard in place of the reasonable care standard.[24] The customary care standard requires juries to limit their inquiry to whether the professional deviated from customary practices in the field. Unlike in ordinary negligence cases, juries in professional malpractice cases cannot be asked to second-guess the reasonableness of the customs in question.[25] Several derivative rules expose additional differences. For example, evidence of customary practices typically must be established by expert testimony, unless the matter is one of common knowledge.[26] Customary practices can include minority "schools of thought" disfavored by a majority of practitioners, allowing for a more heterogeneous set of accepted practices.[27] And, in many states, malpractice claims cannot be heard unless accompanied by a "certificate of merit" signed by a

23. *See id*.

24. *See generally* Dan B. Dobbs et al., Hornbook on Torts § 21.6, at 506 (2d ed. 2016).

25. *See* Tim Cramm et al., *Ascertaining Customary Care in Malpractice Cases: Asking Those Who Know*, 37 Wake Forest L. Rev. 699, 702–03 (2002).

26. *See* Alex Stein, *Toward a Theory of Medical Malpractice*, 97 Iowa L. Rev. 1201, 1213–15 (2012).

27. *See* Gary T. Schwartz, *The Beginning and the Possible End of the Rise of Modern American Tort Law*, 26 Ga. L. Rev. 601, 664–65 (1992).

member of the profession.[28] At the same time, malpractice claims cannot be disclaimed by contractual waiver.[29]

I have argued that the switch to the customary care standard is justified when: (1) practitioners must exercise considerable judgment due to inherent uncertainties in the science of the field; (2) bad outcomes are endemic to the practice because of those uncertainties; and (3) the practice serves a socially vital service even when bad outcomes occur.[30] As it happens, the three factors align exactly with the First Restatement of Torts and its analysis of new technologies.[31] Conversely, when those conditions fade, the need for the professional liability framework wanes concomitantly, and courts cease to give deference to that occupation as a "profession."[32]

I also highlighted three basic misconceptions about the doctrine of professional malpractice. First, the professional care standard is not a higher (or lower) standard of care than the ordinary standard of care. Second, although notions of trust are central to the professional malpractice framework, high trust is not a prerequisite for invoking the doctrine. Third, sociologically derived indicia such as formal education, licensing, or codes of ethics are neither sufficient nor necessary to trigger the professional liability regime.

The first fallacy is that professional malpractice is merely a heightened form of ordinary negligence. Because "professionals" are commonly associated with higher education, training, and social prestige, it is easy to assume that "professional negligence" imposes an elevated duty of care upon those who are more competent. After all, the standard of ordinary reasonable care already incorporates relative levels of knowledge and skill. Yet, if professional malpractice simply means that

---

28. *See* Benjamin Grossberg, Comment, *Uniformity, Federalism, and Tort Reform: The* Erie *Implications of Medical Malpractice Certification of Merit Statutes*, 159 U. PA. L. REV. 217, 222–25 (2010).

29. *See* RESTATEMENT (THIRD) OF TORTS: LIAB. FOR ECON. HARM § 4 (AM. L. INST. 2020) (noting that malpractice is a "prominent exception" to contract-based limitations on liability for financial losses); Catherine M. Sharkey, *Can Data Breach Claims Survive the Economic Loss Rule?*, 66 DEPAUL L. REV. 339, 365 (2017); *see also* Choi, *supra* note 22, at 601 n.209.

30. *See* Choi, *supra* note 22, at 614–15.

31. The First Restatement deemed aviation to be an "ultrahazardous" activity because (1) "aeroplanes have not been so perfected as to make them subject to a certainty of control approximating that of which automobiles are capable," (2) "the serious character of harm which an aeroplane out of control is likely to do," and (3) "aviation has not as yet become either a common or an essential means of transportation." *See* RESTATEMENT OF TORTS § 520 cmt. g (AM. L. INST. 1938). My contention is that if such an activity becomes common or essential to society, but the risks of harm remain serious and continue to elude certainty of control, then a transition to the malpractice framework becomes the most appropriate move.

32. *See* Choi, *supra* note 22, at 618 (describing how courts have "deprofessionalized" architects and engineers by secondguessing expert opinions and applying an ordinary reasonable care standard).

a doctor is held up to the standard of the average doctor having relevant skill and training, then the malpractice doctrine collapses into ordinary negligence and serves no independent purpose.[33]

On the contrary, the customary care standard establishes a self-governance regime checked by judicial oversight. In some aspects, the professional community might embrace expectations that are higher than those a jury might enforce, and in other aspects, those expectations might fall short of what a jury would demand. To be sure, there are valid criticisms of self-governance—a "conspiracy of silence" could allow bad actors to escape accountability.[34] Yet, I have argued that the malpractice doctrine arises when there is no good way to objectively or scientifically evaluate the exercise of professional judgment. It enables courts to impose some accountability without resorting to all-or-nothing measures.

The second fallacy is that a profession must enjoy high levels of societal trust to benefit from the malpractice self-governance regime. But the history of the malpractice doctrine shows that it was created at a time when professions such as medicine and law were relative backwaters of prestige and weakened by infighting.[35] The customary care standard is not a reward for good behavior. It is a product of shortfalls of the reasonable care standard. When doctors lose patients or lawyers lose cases, it is too easy for jurors to find fault, even though such bad outcomes are statistical inevitabilities. In other words, the focus is on the nature of the *work*, not of the professional. The malpractice doctrine is needed when the work performed by professionals is especially likely to lead to unfair second-guessing of professional judgment.

Trust is an important component of the malpractice doctrine, which has evolved to require duties of loyalty in addition to duties of care.[36] But *high* trust is at best an output of the professional malpractice regime, not an input. Instead, the motivation for the doctrine is that public trust will be too low to sustain an otherwise essential service.

---

33. Indeed, numerous commentators over the years have suggested doing away with the professional malpractice doctrine for precisely this reason. *See, e.g.*, DOBBS ET AL., *supra* note 24, at 507; Philip G. Peters, Jr., *The Quiet Demise of Deference to Custom: Malpractice Law at the Millennium*, 57 WASH. & LEE L. REV. 163, 201 (2000) (arguing that the professional paradigm is weakening and that "the custom-based standard of care gradually is yielding to the fundamental tort standard of reasonable care under the circumstances").

34. *Cf.* Wendy Wagner, *When a Corporation's Deliberate Ignorance Causes Harm: Charting a New Role for Tort Law*, 72 DEPAUL L. REV. 413 (2022) (criticizing corporations for manipulating the scientific record to downplay the hazardousness of their activities).

35. *See generally* WILLIAM G. ROTHSTEIN, AMERICAN PHYSICIANS IN THE 19TH CENTURY: FROM SECTS TO SCIENCE (1972); PAUL STARR, THE SOCIAL TRANSFORMATION OF AMERICANA MEDICINE (1982).

36. *See* Choi, *supra* note 22, at 609; *cf.* Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961 (2021) (proposing a duty of loyalty for data collectors and tech companies).

When bad outcomes are a statistical inevitability, and the practice of the profession is an inexact art rather than a precise science, then it is too easy to lose faith and to condemn such services.

The third common fallacy is that professions are defined by personal traits such as education, salary, licensure, or "gentlemanly" culture.[37] Yet, for prototypical professions such as medicine and law, this theory is historically implausible, given the protean state of professional education and professional ethics at the time when the malpractice doctrine was first applied and extended. Nor does it adequately explain, as a matter of positive law, the continued exclusion of many occupations that have strived to meet precisely those criteria. Finally, on a normative level, the malpractice rule should not function as a special rule for elites. The reasonable care standard is appropriate for most lines of work, even those that are highly compensated, highly specialized, and highly ethical. Conversely, the *absence* of characteristics such as formal education requirements, licensure schemes, disciplinary systems, or ethical codes should not be a valid reason to bar entry into the malpractice regime where it is otherwise appropriate.

II. Matters of Professional Judgment

The first factor relevant to the legal designation of a profession is whether the core elements of the work involve substantial uncertainties in knowledge, and therefore require latitude for discretionary judgment. On the surface, both AI and software seem alike in that each involves code and data, in a field of rapid innovation where most practices are ad hoc and experimental rather than established through scientific method. Nevertheless, the uncertainties that arise in software work appear to be more fundamental and pervasive to the enterprise, whereas those that arise in AI work are quite different in nature and arguably limited in scope.

The most material distinction between AI work and conventional software work is the bottom-up versus top-down process by which such systems are constructed. Whereas software work amplifies and celebrates complexity in ways that defy expert understanding, AI work seeks to simplify a complex pattern and encapsulate it within a unified

37. *See* Choi, *supra* note 22, at 589 (citing Hosp. Comput. Sys., Inc. v. Staten Island Hosp., 788 F. Supp. 1351, 1361 (D.N.J. 1992)). More recently, the Third Restatement of Torts has embraced this list of traits and defended them on policy grounds that the professional-client relationship entails unique risks that cannot be effectively regulated by contract law. *See* Restatement (Third) of Torts: Liab. for Econ. Harm § 4 cmt. b (Am. L. Inst. 2020). Yet, this definition of professional as risk-bearer is circular and also inconsistent with other aspects of the professional malpractice doctrine. *See* Choi, *supra* note 22, at 611 (critiquing the careless merger of professional duties and fiduciary duties).

mathematical model. That fundamental difference offers greater hope that much of AI work could be conducive to standardization, which has eluded software work.

Conventional software is constructed in a top-down manner in the sense that every line of code is written for a command flow purpose. The developer team determines the system requirements, converts those requirements into a preliminary design, implements a working model in code, and tests the code to ensure it performs as specified. That is not to say the software development process is centralized or monolithic. Modern best practices encourage software developers to work in fluid, decentralized phases through methods such as iterative cycles and "agile" methods. Nevertheless, all conventional software systems are ultimately composed of individual lines of code that are stacked together with intentional, human-designed control paths in mind.

By contrast, the construction of modern AI systems can be characterized as bottom-up because their architectures are governed primarily by data-led patterns, rather than by human-led blueprints.[38] The core component of any AI system is its knowledge representation model, which enables it to incorporate prior experience and expertise into its decisions. For example, a chess program is more likely to win if it knows the best possible moves, and a self-driving car is less likely to crash if it understands common road signals and obstacles. In modern, real-world AI systems, these knowledge representation models are not configured by hand, but automatically configured using "machine learning" techniques that integrate large datasets of training examples.[39] If the data were truly random, it would be impossible to process so much data; instead, the AI modeler assumes that most meaningful data from the real world will follow a naturally sparse pattern. The purpose of any learning algorithm is to sculpt a model that approximates as well as possible that naturally existing pattern.

That difference between AI and conventional software was not always so stark. "Classical" approaches to AI draw on rules of formal logic to

---

38. *See* Harry Surden, *Machine Learning and Law*, 89 Wash. L. Rev. 87, 93–94 (2014); *cf.* Andrej Karpathy, *Software 2.0*, Medium (Nov. 11, 2017) ("It turns out that a large portion of real-world problems have the property that it is significantly easier to collect the data (or more generally, identify a desirable behavior) than to explicitly write the program.").

39. *See* Ian Goodfellow, Yoshua Bengio & Aaron Courville, Deep Learning 96 (MIT Press 2016) ("A machine learning algorithm is an algorithm that is able to learn from data."); *id.* at 8 ("We contend that machine learning is the only viable approach to building AI systems that can operate in complicated real-world environments."); Joel Klinger et al., A Narrowing of AI Research? (Jan. 11, 2022) (preprint), https://arxiv.org/pdf/2009.10385.pdf [https://perma.cc/V4CX-VZZP] (finding that AI research in deep learning techniques has expanded rapidly while AI research in classical methods—such as symbolic representation and statistical machine learning—has stagnated, particularly in the commercial sector).

manually encode the knowledge base needed for a task—an approach that is functionally indistinguishable from conventional software.[40] For example, to tackle a problem like medical diagnosis of blood infections, the AI modeler might seek to identify all possible symptoms and diagnoses, and then manually map all possible relationships between those symptoms and diagnoses.[41] When those rules are written out by hand, they are, in fact, lines of code. In other words, classical AI systems were designed and implemented in the same top-down manner as conventional software systems. Although defenders of the classical rule-based approaches maintain that true semantic knowledge cannot be codified in any other way,[42] none has achieved meaningful success in real-world applications.[43]

Instead, modern AI techniques have surpassed the limitations of classical AI approaches by improving methods that auto-generate knowledge representation models. The key innovation has been the discovery that one particular model structure—deep neural networks—can be trained at sufficient size[44] and capacity[45] to perform well on real-world

40. *See* Marta Garnelo & Murray Shanahan, *Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations*, Current Opinion in Behavioral Scis., Oct. 2019, at 17, 17 (describing symbolic AI as "handcrafted" rather than "learned from data").

41. *See* Stuart Russell & Peter Norvig, Artificial Intelligence: A Modern Approach 23 (3d ed. 2010) (describing early expert systems such as the Mycin system for diagnosing blood infections, which incorporated knowledge "acquired from extensive interviewing of experts").

42. *See, e.g.*, Gary Marcus & Ernest Davis, *GPT-3, Bloviator: OpenAI's Language Generator Has No Idea What It's Talking About*, MIT Tech. Rev. (Aug. 22, 2020), https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/ [https://perma.cc/V623-CLGC].

43. *See* Goodfellow et al., *supra* note 39, at 2 (observing that several AI projects "have sought to hard-code knowledge about the world in formal languages," but that "[n]one of these projects has led to a major success" because it is difficult "to devise formal rules with enough complexity to accurately describe the world"); Russell & Norvig, *supra* note 41, at 439 ("[T]he enterprise of general ontological engineering has so far had only limited success. None of the top AI applications . . . make use of a shared ontology—they all use special-purpose knowledge engineering."); *see also* Calo, *supra* note 12, at 404–05 (describing the failure of symbolic systems to "yield many viable applications in practice," which led to "the dwindling of research funding in the late 1980s known as the 'AI Winter.'").

44. *See* Goodfellow et al., *supra* note 39, at 431 ("One of the key factors responsible for the improvement in neural network's accuracy and the improvement of the complexity of tasks they can solve between the 1980s and today is the dramatic increase in the size of the networks we use.").

45. The richness and expressive power of a model is called its representational "capacity." *See* Vladimir Vapnik, Esther Levin & Yann Le Cun, *Measuring the VC-dimension of a Learning Machine*, 6 Neural Computation 851, 851–52 (1994) (defining the "capacity" of a learning machine as the complete set of classification functions from which an optimal solution can be chosen); Yaser S. Abu-Mostafa, *The Vapnik-Chervonenkis Dimension: Information Versus Complexity in Learning*, 1 Neural Computation 312, 316 (1989) (explaining that a neural network's capacity is correlated with its size).

tasks.[46] As early researchers showed, an AI system performs best when its model capacity matches the true complexity of the real-world task.[47] The ability to construct and optimize multiple intermediate layers has been critical to the success of neural networks, because depth can augment the capacity of the model by exponential orders of magnitude.[48]

In a basic neural network, each neuron node unit represents a single monad of information. Multiple node units are stitched together to form a layer, with different "weights" assigned to each connection. As the input data travels through each layer, the weights determine which nodes are activated, which directs a path through the deep neural network to produce the output result. Multiple layers are then stacked sequentially so that each layer feeds forward into the next layer. That multilayer structure enables the AI modeler to represent very complex concepts in a more efficient manner. For example, an image of a person could be composed of individual pixels at the input layer; edges and contours at the next layers; higher-order elements such as eyes and lips at successive layers; and so on until ultimately a facial identification is presented at the output layer.

The structure of a deep neural network model is configured automatically by the learning algorithm and by the data, through a process called backpropagation (or "backprop"), a technique originally introduced in 1986.[49] All leading deep learning libraries, including PyTorch and TensorFlow, are built upon the backprop method.[50] As the backprop algorithm iterates through the data, individual connections between neuron node units are reinforced or decayed, so that the model gradually learns to recognize patterns it has seen before and to disregard patterns that are absent in the dataset. Like a toddler learning language, unused

---

46. *See* Goodfellow et al., *supra* note 39, at 20–21 ("As of 2016, a rough rule of thumb is that a supervised deep learning algorithm . . . will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples.").

47. *See id.* at 109 ("Machine learning algorithms will generally perform best when their capacity is appropriate for the true complexity of the task they need to perform and the amount of training data they are provided with. Models with insufficient capacity are unable to solve complex tasks. . . . [B]ut when their capacity is higher than needed to solve the present task, they may overfit.").

48. The reason for the advantage generated by deep networks remains undertheorized. However, scholars have suggested that it relates to the superior ability of deep networks to approximate "compositional functions," i.e., functions that are composed of a hierarchy of constituent functions. *See* Tomaso Poggio et al., *Why and When Can Deep—but Not Shallow—Networks Avoid the Curse of Dimensionality: A Review*, 14 Int'l J. Automation & Computing 503, 503 (2017).

49. *See* Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 Harv. J.L. & Tech. 889, 901 n.50 (2018) (describing history of the backpropagation algorithm).

50. *See* Nathan Sprague, *ScalarFlow: Implementing Reverse Mode Automatic Differentiation*, AI Matters, Dec. 2021, at 8, https://sigai.acm.org/static/aimatters/7-4/AIMatters-7-4-04-Sprague.pdf [https://perma.cc/9D83-478Y].

neural connections may be pruned to create a sparser network that can be traversed more quickly and efficiently.

At the outset, the AI modeler is responsible for selecting the model structure, the learning algorithms, and certain "hyperparameters" that govern the overall learning process. The AI modeler also needs to collect and preprocess the training data. But once the learning process begins, the AI model is auto-generated without further intervention by the AI modeler. If the resulting model is unsatisfactory, the AI modeler can generate a new model by adjusting the initial configuration settings or by refining the dataset. The search for a satisfactory model is typically tedious and unpredictable. But the amount of conventional software code that needs to be rewritten is minimal.

This bottom-up construction of AI systems differs from top-down construction of conventional software systems in at least three salient ways. First, the amount of software design involved is minimal. Instead, AI modelers typically use off-the-shelf software and adjust only a handful of initialization settings and hyperparameters. Second, the training data plays an outsized role in determining how the AI system behaves. Yet, while the creation of new datasets is tedious and time-consuming, that work is largely rote and perfunctory. Moreover, it is common practice to outsource such work or to reuse preexisting datasets compiled by others. Third, error metrics are more easily quantifiable than the error metrics available for conventional software systems. That said, those error metrics have important limitations. The remainder of this Section proceeds in greater detail through each of the three aspects.

A. Learning Algorithms and Hyperparameters

The AI modeler must make several initialization choices at the outset. In theory, the range of available configurations is quite large, but in practice, those choices are generally confined to what is known to have worked well before. When AI experts claim that the work of training deep neural networks is more an "art" than a "science," it is because many of these initialization choices are determined by trial and error, rather than by precise knowledge. Nevertheless, it can be argued that the overall range of judgment being exercised by the average practitioner is fairly narrow. Moreover, if the AI modeler is retraining a preexisting AI model, then many choices are already locked in.[51]

First, the network structure must be optimized for the type of learning task. For smaller datasets, "convolutional" models are favored for

51. *See* Edward J. Hu et al., *LoRA: Low-Rank Adaptation of Large Language Models*, 10 Int'l Conf. on Learning Representations (2022).

image-based classifications, while "recurrent" models are preferred for natural language processing tasks. In the past few years, newly developed techniques in unsupervised learning have shifted momentum to the "transformer" model, which enables processing of much more massive amounts of data, and which forms the basis of large language models such as ChatGPT.[52] These architectural choices are driven primarily by evidence of prior experimental success.

Second, in determining the network structure, the AI modeler also has the option of choosing the depth of the network, the number of nodes (width) per layer, and other "hyperparameters" that govern the overall learning process. The length of the longest path from input to output defines the "depth" of the neural network.[53] To be clear, a deeper network is not always a superior one; the best depth is one that matches the true complexity of the task. For smaller datasets, one or two hidden layers is usually sufficient to yield best results. It is preferable to make each layer sufficiently wide by adding more nodes per layer than it is to add further depth.[54] When the available training datasets are more massive, however, and the learning problem more complex, then deeper networks offer the potential to generate better results.[55]

Next, the AI modeler must choose the learning algorithm, which consists of mathematical "activation functions," a "loss function," and an "optimization technique." The goal is to learn the best node weights across the entire neural network, which can include millions or billions of nodes. In lay terms, (1) an activation function is what instructs each neuron node to fire or not to fire, (2) the loss function represents the penalty for giving wrong weights to those nodes, while (3) the optimization technique is the iterative learning strategy that adjusts node weights to minimize the penalty of being wrong. With proper choices, the node weights will converge toward the optimal configuration.

The potential universe of activation functions and loss functions is vast, but in practice, there is a narrow set of commonly used options, which are further limited by the type of learning problem and the data

---

52. *See* Ashish Vaswani et al., *Attention Is All You Need*, 30 Advances Neural Info. Processing Sys. (2017) https://arxiv.org/pdf/1706.03762.pdf [https://perma.cc/9FM9-SGBL].

53. *See* Goodfellow et al., *supra* note 39, at 7–8 (defining depth as "the length of the longest path from input to output" but cautioning that "there is no single correct value for the depth of an architecture," and that there is no "consensus about how much depth a model requires to qualify as 'deep'").

54. *See* Zeyuan Allen-Zhu et al., *Learning and Generalization in Overparameterized Neural Networks, Going Beyond Two Layers*, 32 Advances Neural Info. Processing Sys. 6158 (2019).

55. *See* Rupesh Kumar Srivastava et al., *Training Very Deep Networks*, 28 Advances Neural Info. Processing Sys. 2377, 2377 (2015); Kaiming He et al., *Deep Residual Learning for Image Recognition*, 2016 IEEE Conf. on Comput. Vision & Pattern Recognition 770, 770 (describing novel "residual learning" technique to achieve 152 layers of depth).

distribution.[56] The choice is even more stark for optimization techniques, where the most popular choice is stochastic gradient descent, or improved versions thereof. Within this gradient descent paradigm, the AI modeler must choose another set of hyperparameters to determine how crudely or finely the stochastic gradient descent algorithm explores the universe of possible node weight configurations. For example, variables like learning rate, batch size, momentum, and weight decay are used to determine how much to change the node weights after each iteration of the training process.

These hyperparameters regulate how much and how quickly the algorithm learns from the data. If the algorithm learns too slowly, it suffers from *underfitting*, which means it fails to learn the task at hand. Conversely, if the algorithm learns too quickly, it can suffer from *overfitting*, which is a problem because the resulting model performs deceptively well on the training examples but fails to generalize well to new examples.[57] In order to avoid the twin hazards of overfitting and underfitting, one must choose appropriate hyperparameters that lead to a well-balanced model.

If there is a dark art to training deep neural networks, it is in the selection of these mathematical functions and hyperparameters.[58] To be sure, there is no current theory that allows AI modelers to predict with precision which values will work best prior to training. And, because AI training is a computationally intensive process, it is prohibitively expensive to brute force through all possible choices.

---

56. For example, PyTorch offers 30 built-in activation functions and 21 built-in loss functions. *See Torch.nn*, PyTorch, https://pytorch.org/docs/stable/nn.html [https://perma.cc/PL6S-AM6W]. TensorFlow offers 17 built-in activation functions and 19 classes of built-in loss functions. *See Module: tf.keras.activations*, TensorFlow, https://www.tensorflow.org/api_docs/python/tf/keras/activations [https://perma.cc/NJ28-XL3V]; *Module: tf.keras.losses*, TensorFlow, https://www.tensorflow.org/api_docs/python/tf/keras/losses [https://perma.cc/E2P6-YBFL]. The AI modeler can also implement customized activation functions or loss functions using both libraries. For loss functions, one can choose a classification loss function or a regression loss function. Classification functions work best when identifying a set of discrete values (e.g., the numerical digits 0 to 9). Regression functions are necessary when seeking to learn values that are on a smooth, continuous spectrum. Within each of these two categories, there is a short list of mathematical functions that are known to work well for deep learning applications.

57. *See* Goodfellow et al., *supra* note 39, at 108 ("Underfitting occurs when the model is not able to obtain a sufficiently low error value on the training set. Overfitting occurs when the gap between the training error and test error is too large.").

58. *See* James Bergstra et al., *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*, 30 Int'l Conf. on Machine Learning 115, 115 (2013) (observing that the tuning process "often depends on personal experience and intuition in ways that are hard to quantify or describe"); Goodfellow et al., *supra* note 39, at 420 ("Manual hyperparameter tuning can work very well when the user has a good starting point, such as one determined by others having worked on the same type of application and architecture, or when the user has months or years of experience in exploring hyperparameter values for neural networks applied to similar tasks.").

On the other hand, research suggests that the range of relevant choices is relatively small.[59] And more advanced techniques are allowing AI modelers to automatically tune hyperparameters with increasing ease, rather than relying on manually selected defaults.[60] To be sure, there is no hard and fast rule stopping anyone from making wildly different choices, but there are general expectations among experts in the field as to the ranges of values that tend to work well.

B. Training Data

It is axiomatic that the simplest way to improve the performance of a deep neural network is to increase the size of the training dataset.[61] Thus, much of the human labor that goes into building deep learning AI models is collecting and curating data.[62] David Lehr and Paul Ohm have argued that legal scholars need to "giv[e] more attention to machine learning's playing-with-the-data stages."[63] Likewise, Andrew Selbst and Solon Barocas have also stated that it is necessary to investigate the process behind a model's development, not just the running model itself.[64] Frank Pasquale has advocated the need to impose duties of care on firms that rely on faulty data to develop AI models.[65] Having said that, much of the work at the data supervision stage is mechanical and tedious, and therefore delegated to low-skilled workers. Alternatively, many AI modelers reuse existing datasets rather than generate entirely new datasets. Where AI modelers may bear the brunt of responsibility

---

59. *See* James Bergstra & Yoshua Bengio, *Random Search for Hyper-Parameter Optimization*, 13 J. Machine Learning Rsch. 281 (2012) (finding that most hyperparameters do not matter much for most data sets). *But see* Goodfellow et al., *supra* note 39, at 420 ("Neural networks can sometimes perform well with only a small number of tuned hyperparameters, but often benefit significantly from tuning of forty or more.").

60. *See, e.g.*, Greg Yang et al., *Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer*, 34 Advances Neural Info. Processing Sys. (2021) (describing efficient methodology to tune hyperparameters on a scaled-down model and then transfer those values to the full-scale model). *But see* Goodfellow et al., *supra* note 39, at 424 (asserting, at the time of the writing, that hyperparameter optimization is much less efficient than manual search by a human practitioner).

61. *See* Goodfellow et al., *supra* note 39, at 414 ("Many machine learning novices are tempted to make improvements by trying out many different algorithms. Yet, it is often much better to gather more data than to improve the learning algorithm.").

62. *See* Lehr & Ohm, *supra* note 18, at 677 ("For many projects, [data collection] can be the most time-consuming stage, and it also holds enormous consequences; as commenters have noted previously, an algorithm is, at the end of the day, only as good as its data.").

63. *Id.* at 656 ("The potential harms and benefits that can creep in while playing with the data differ from those of the running model.").

64. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 Fordham L. Rev. 1085 (2018).

65. *See* Frank Pasquale, *Data-Informed Duties in AI Development*, 119 Colum. L. Rev. 1917 (2019).

for data-related errors is in their use of algorithmic techniques to enhance, augment, or manipulate the data in order to extract better learning outcomes from a limited dataset.

To date, most deep learning research has been on "supervised learning" techniques, meaning that the AI modeler must carefully "supervise" the data to ensure that each example is correctly labeled with the "ground truth" before it is presented to the learning algorithm.[66] The emerging sea change toward semisupervised learning is diminishing the need for manual labeling. Nevertheless, there will always be a need to reduce raw data into more digestible forms, and to augment that data to generate more learning examples. This data preprocessing work has two main functions: first, to simplify the learning task by making the data clean and orderly, and second, to improve the generalizability of the AI model so that it can perform as well as possible in new, unforeseen situations.

With regard to this data supervision work, at least two types of errors could be attributed to the AI modeler: *incorrect* data and *insufficient* data. The first is that the supervisory steps may be performed in a faulty manner that causes the training data to be inaccurate or misleading. The second is that the need for data supervision restricts the available supply of training data, which increases the risk of blind spots in the data distribution. Additionally, legal scholarship has emphasized a third type of error: *illegitimate* data.[67] For example, the data could be problematic for reasons such as breach of privacy, discriminatory bias, or violation of copyright.

1. Incorrect Data

Concerns about inaccurate data are age-old.[68] Data records can contain explicit and implicit errors, and thus advocates and lawmakers have

66. *See* Steven A. Israel et al., *Applied Machine Learning Strategies*, IEEE POTENTIALS, May–June 2020, at 38, 38 ("The most prominent ML methods in use today are supervised, meaning they require ground-truth labeling of the data on which they are trained."). While unsupervised and semi-supervised learning techniques are areas of active research, they are not yet mainstream tools the way that supervised learning techniques have become. Moreover, due to computational limits, even unsupervised and semi-supervised learning techniques need to reduce raw data into more digestible forms.

67. *See* Pasquale, *supra* note 65, at 1923–28 (contrasting "inappropriate data" with "inaccurate data").

68. *See, e.g.*, DANIEL J. SOLOVE, THE DIGITAL PERSON 15 (2004) (quoting a 1973 report by the U.S. Department of Health, Education, and Welfare: "Sometimes the individual does not even know that an organization maintains a record about him. Often he may not see it, much less contest its accuracy, control its dissemination, or challenge its use by others."); *id.* at 46, 49 ("Not only are our digital biographies reductive, but they are often inaccurate. . . . [T]he information in databases often fails to capture the texture of our lives. Rather than provide a nuanced portrait of our

long demanded procedural protections such as rights of public access to data records, and remedial mechanisms to dispute the information found therein.[69]

In order to prepare a training dataset for a supervised learning algorithm, the AI modeler must perform several idiosyncratic steps that are not practiced in conventional software development, including *labeling* and *cleaning* the training data. Each of these intermediate steps has a significant impact on the performance of the ensuing AI model, and each entails varying degrees of uncertainty.

Data labeling is a basic way in which errors can corrupt the machine learning process. For example, if the goal were to distinguish cat images from dog images, then labeling each image in the dataset as "cat," "dog," or some other value teaches the AI model how to correctly classify each image.[70] Likewise, in the autonomous driving context, bounding boxes can be used to label objects of interest such as vehicles, pedestrians, and other road users, which the learning algorithm then knows to avoid. Without that supervisory help, the learning algorithm would need to learn those labels on its own.[71]

But data labeling is not an exact science.[72] Some judgment is needed in designating a useful schema. A naive approach could ask labelers to invent their own labels.[73] Given enough labelers, however, this approach leads to inconsistent use of labels. Additionally, the quality of labels suffers because humans tend to label objects at a basic semantic level rather than in more descriptively rich ways.[74] The labeling process can

---

personalities, compilations of data capture the brute facts of what we do without the reasons."); Simson Garfinkel, Database Nation (2000).

69. *See, e.g.*, Fair Credit Reporting Act, Pub. L. No. 91-508, 84 Stat. 1114, 1127 (1970) (codified at 15 U.S.C. § 1681); Freedom of Information Act, Pub. L. No. 90-23, 81 Stat. 54 (1967) (codified at 5 U.S.C. § 552); California Consumer Privacy Act of 2018, Cal. Civ. Code §§ 1798.100–.199 (West 2022).

70. Labels also correspond to the "outcome variables" that the AI model is expected to predict or estimate. *See* Lehr & Ohm, *supra* note 18, at 673.

71. *See* Goodfellow et al., *supra* note 39, at 102 ("The term supervised learning originates from the view of the target y being provided by an instructor or teacher who shows the machine learning system what to do. In unsupervised learning, there is no instructor or teacher, and the algorithm must learn to make sense of the data without this guide.").

72. *See* Lehr & Ohm, *supra* note 18, at 673–75 (describing the discretionary process by which AI modelers choose labels (or outcome variables), and listing the "different factors [that] play into how data scientists make these tough choices," including subject matter knowledge, algorithmic needs, and resource constraints).

73. In practice, annotations are performed by hand. *See, e.g.*, Jia Deng et al., *ImageNet: A Large-Scale Hierarchical Image Database*, 2009 IEEE Conf. on Comput. Vision & Pattern Recognition 248, 251 ("To collect a highly accurate dataset, we rely on humans to verify each candidate image collected . . . ."); Luis von Ahn & Laura Dabbish, *Labeling Images with a Computer Game*, 2004 SIGCHI Conf. on Human Factors Computing Sys. 319, 319.

74. *See* Deng et al., *supra* note 73, at 250 (explaining findings that "humans tend to label visual objects at an easily accessible semantic level termed as 'basic level' (e.g. bird), as opposed to more

also be sabotaged on purpose.[75] Even small changes to data labels can cause surprisingly nuanced problems that are difficult to detect.

One workaround is to adopt an existing schema developed by linguistic experts or subject matter experts. Perhaps the best known example is ImageNet, an image dataset of over 14 million images organized according to a hierarchy of more than 100,000 labels.[76] The key to its success is that it starts with a third-party schema and then asks the public to provide examples that fit the labels, rather than vice versa.[77] Other research efforts include methods to automate data labeling,[78] which have yielded limited success so far.[79]

Data cleaning allows the AI modeler to massage the data to yield better results.[80] If the data is too noisy, then the learning algorithm will struggle to distinguish features that are meaningful from those that are not.[81] The dataset may contain faulty or inappropriate entries,[82] or the

---

specific level ('sub-ordinate level', e.g. sparrow), or more general level ('super-ordinate level', e.g. vertebrate).").

75. *See* Antonio Torralba et al., Open Letter, MIT (June 29, 2020), https://groups.csail.mit.edu/vision/TinyImages/ [https://perma.cc/ELL3-6DZN] (announcing formal withdrawal of the Tiny Images dataset due to the discovery of "derogatory" labels and "offensive images").

76. *About ImageNet*, IMAGENET, http://image-net.org/about.php [https://perma.cc/7FAF-BKVT]; Deng et al., *supra* note 73, at 248; *see also* Abeba Birhane & Vinay Uday Prabhu, *Large Image Datasets: A Pyrrhic Win for Computer Vision?*, 2021 IEEE WINTER CONF. ON APPLICATIONS COMPUT. VISION 1536, 1539 ("ImageNet, with its vast amount of data, has not only erected a canonical landmark in the history of AI, it has also paved the way for even bigger, more powerful, and suspiciously opaque datasets."). *But see* Kate Crawford & Trevor Paglen, *Excavating AI: The Politics of Images in Machine Learning Training Sets*, AI NOW INST. (2019), https://excavating.ai [https://perma.cc/6FGS-T3DK] (detailing at length the shortfalls of the ImageNet taxonomy).

77. *See* Deng et al., *supra* note 73, at 248, 251–52.

78. *See* Burr Settles, *Active Learning Literature Survey*, 1648 U. WIS. COMPUT. SCIS. TECH. REP. (2009), at 9, https://minds.wisconsin.edu/bitstream/handle/1793/60660/TR1648.pdf [https://perma.cc/752Q-YBTX].

79. *See* GOODFELLOW ET AL., *supra* note 39, at 526 ("Today, unsupervised pretraining has been largely abandoned, except in the field of natural language processing . . . ."). *But see* Alec Radford et al., Improving Language Understanding by Generative Pre-Training (June 11, 2018) (preprint), http://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [https://perma.cc/DB7S-WRSJ] (explaining use of unsupervised pretraining techniques to train GPT-2 and GPT-3 on large unlabeled text corpora scraped from the internet).

80. *See* IHAB F. ILYAS & XU CHU, DATA CLEANING 1 (2019) ("[D]ata cleaning activities usually consist of two phases: (1) error detection, where various errors and violations are identified and possibly validated by experts; and (2) error repair, where updates to the database are applied (or suggested by human experts) to bring the data to a cleaner state suitable for downstream applications and analytics.").

81. *See* GOODFELLOW ET AL., *supra* note 39, at 414 ("If large models and carefully tuned optimization algorithms do not work well, then the problem might be the *quality* of the training data. The data may be too noisy or may not include the right inputs needed to predict the desired outputs. This suggests starting over, collecting cleaner data, or collecting a richer set of features.").

82. Such errors can include, for example, missing values, unhelpful outliers, typos, and duplicate entries.

data may need to be normalized to share the same units and measures.[83] With enough data, small levels of error can become invisible, but systematic errors will continue to be problematic. For certain AI applications, these data cleaning tasks may require domain-specific knowledge, but for "common sense" applications, they may require little to no specialized skill.

Another way to enhance the data is through "feature engineering," which allows the AI modeler to highlight or diminish certain features of the training dataset. As an example, one popular technique is pooling, which combines several existing features into a new aggregate feature.[84] For example, if the dataset contains hourly wage information, the AI modeler could pool that information to create a new weekly wage feature. Likewise for image processing, a pooling function can produce one aggregate value for multiple neighboring pixels. In this way, the AI modeler simplifies the learning task while still preserving the general contours of the original dataset.

In sum, any exercise of data curation has the potential to introduce bias or error, and choosing the best techniques involves some guesswork. Nevertheless, the ongoing development of standard schemas and certified datasets could reduce much of that uncertainty going forward.

2. Insufficient Data

The problem of insufficient data is a fundamental challenge for all machine learning methods, because it increases the likelihood that the trained AI model will fail to generalize well to new, previously unseen cases. When a new case reveals a gap in the data, the AI model must rely on inferential reasoning, which creates uncertainty and risk of error. Conversely, as the amount of available data increases to infinity, the more likely it becomes that the AI model will have seen the full range of cases, thus eliminating the need for guesswork. The sufficiency concern is not merely quantitative but qualitative. Facial recognition systems have been criticized for performing worse on darker faces.[85] Medical

83. *See* GOODFELLOW ET AL., *supra* note 39, at 517 ("Many information processing tasks can be very easy or very difficult depending on how the information is represented . . . . For example, it is straightforward for a person to divide 210 by 6 using long division. . . . Most modern people asked to divide CCX by VI would begin by converting the [Roman] numbers to the Arabic numeral representation . . . ."); *id.* at 441 ("Many computer vision architectures require images of a standard size, so images must be cropped or scaled to fit that size.").

84. *See* Naila Murray & Florent Perronnin, *Generalized Max Pooling*, 2014 IEEE CONF. ON COMPUT. VISION & PATTERN RECOGNITION 2473.

85. *See* Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, 81 PROC. MACH. LEARNING RSCH. 1, 8 (2018).

datasets often display gender gaps that overlook women.[86] Language models have come under attack for "hallucinating" false facts.[87] These faults occur because the training data is not adequately representative of the full range of information that exists in the real world.

The amount of data needed to achieve passable performance depends on the complexity of the problem. If there are only a few causal factors that affect the outcome—such as a lamp controlled by a simple on-off switch—then even a small number of past examples will be strongly predictive of future cases. However, most real-world applications of AI involve problems that are not well-understood. If the problem is more like weather forecasting, where outcomes can vary dramatically depending on a vast set of unknown variables, then much more data is needed to record the myriad possible patterns. One benchmark suggested in the literature is that 10 million labeled examples is the bare minimum number needed to achieve human-like performance.[88] But the critical threshold can be much higher. For example, the autonomous car company Waymo states that it has turned its "20 million miles of on-road experience into a searchable catalog of billions of objects."[89] OpenAI's GPT-3 (the predecessor of ChatGPT) was trained on 499 billion tokens.[90]

Actual practice among AI modelers varies considerably. As the foregoing discussion emphasizes, the need to label data drastically limits the availability of training data. Other compounding factors—such as data privacy laws—can also inhibit the ready accessibility of usable data. To offset this shortage of training data, AI modelers often rely on "found" data.[91]

---

86. *See* Caroline Criado Perez, Invisible Women: Data Bias in a World Designed for Men 89–90 (2019).

87. *See* Ziwei Ji et al., *Survey of Hallucination in Natural Language Generation*, 55 ACM Computing Surveys art. 248, at 248:5–248:6 (2023) (observing that the main causes of hallucination include problems with data collection, as well as data distortions that can arise through training and modeling choices).

88. *See* Goodfellow et al., *supra* note 39, at 20.

89. *See* James Guo et al., *Seeing is Knowing: Advances in Search and Image Recognition Train Waymo's Self-Driving Technology for Any Encounter*, Waypoint (Feb. 6, 2020), https://blog.waymo.com/2020/02/content-search.html [https://perma.cc/Z8C5-8QDQ]; *see also* Pei Sun et al., *Scalability in Perception for Autonomous Driving: Waymo Open Dataset*, 2020 IEEE Conf. on Comput. Vision & Pattern Recognition 2443, 2443–44.

90. *See* Tom B. Brown et al., *Language Models Are Few-Shot Learners*, 33 Advances Neural Info. Processing Sys. 1877 (2020) (describing implementation of GPT-3).

91. *See* Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 Colum. L. Rev. 1851, 1861 (2019) ("[M]achine learning processes often rely on 'found data,' collected for some other purpose, to train the models. Unfortunately, reliance on found data leaves rulemakers at the mercy of whatever feature sets and outcome variables happen to have been collected.").

When the available dataset is too small, the AI modeler must get creative in order to extract more learning from less data. Some techniques allow for careful reuse of the existing data.[92] More popular techniques, however, use data augmentation to generate new examples. The AI modeler can greatly magnify the size of a dataset by adding slight perturbations of existing examples in the dataset. For example, an image can be rotated, flipped, or altered in minor ways that preserve the essence of the image while making the image appear new and different to the learning algorithm.[93] In language corpuses, synonyms can be substituted to create new text examples with equivalent meaning. More advanced techniques for generating new data samples include diffusion and generative adversarial networks.[94] By massaging the available data in these ways, AI modelers boost the amount of learning that can be extracted from limited quantities of training data. But doing so greatly enhances the risk of error if the augmentation techniques are not carefully implemented.[95]

3. Illegitimate Data

A third source of potential fault focuses on whether certain data in the training dataset is illegitimate.[96] Here, there are two genres of critique. The strong form is that the data is *malum in se* and cannot be purged of its problematic aspects. The weaker version raises procedural objections regarding the manner in which the data was obtained or processed.

One set of objections stems from the intimate or offensive nature of the data itself.[97] For example, facial recognition systems have generated

---

92. For example, one clever way to extract more learning from a single dataset is to divide the dataset into *k* nonoverlapping subsets, and then repeat the training and testing procedures *k* times. This technique is known as "*k*-fold cross-validation." *See* GOODFELLOW ET AL., *supra* note 39, at 119.

93. *See* Alex Ratner et al., *Learning to Compose Domain-Specific Transformations for Data Augmentation*, STAN. DAWN (Aug. 30, 2017), https://dawn.cs.stanford.edu/2017/08/30/tanda [https://perma.cc/8T8H-XZYE].

94. *See* Prafulla Dhariwal & Alex Nichol, Diffusion Models Beat GANs on Image Synthesis (June 1, 2021) (preprint), https://arxiv.org/pdf/2105.05233.pdf [https://perma.cc/42KW-HKCJ]. Other methods such as fine tuning and transfer learning offer additional ways to extract utility from smaller datasets. *See* Rohan Taori et al., *Alpaca: A Strong, Replicable Instruction-Following Model*, STAN. UNIV. CTR. RSCH. ON FOUND. MODELS (2023), https://crfm.stanford.edu/2023/03/13/alpaca.html [https://perma.cc/6JCQ-F9ZW].

95. *See* Sina Alemohammad et al., Self-Consuming Generative Models Go MAD (July 4, 2023) (preprint), https://arxiv.org/pdf/2307.01850v1.pdf [https://perma.cc/DZN8-U9VL] (finding that overuse of synthetic, AI-generated training data degrades the quality of future AI models).

96. *See* Pasquale, *supra* note 65, at 1925–27.

97. *See* Karen E.C. Levy, *Intimate Surveillance*, 51 IDAHO L. REV. 679 (2015); Margaret Hu, *Biometric ID Cybersurveillance*, 88 IND. L.J. 1475 (2013).

public outcry and have led to bans by multiple municipalities.[98] The use of deep learning for biometric recognition is not limited to faces, but extends to fingerprints, palmprints, eyes, voices, gaits, handwriting, and ears.[99] Similarly, the use of nude content to generate "deep fake" pornography raises objections rooted in bodily autonomy and shock to the conscience.[100] This indignation could be extended to other categories of information—health data, children's data, financial data, geolocational data, and so on.

Other concerns have been framed in procedural terms, in which the data is not illegitimate per se, but becomes tarnished by the manner in which it is acquired or used.[101] Complaints rooted in consent,[102] intellectual property,[103] contextual integrity,[104] or anonymization,[105] speak primarily to failures of data handling that could be remediated by better procedural safeguards or economic compensation.

A more radical set of concerns is that reliance on biased, historical data will sustain an unjust status quo.[106] The canonical case study

---

98. *See* Lindsey Barrett, *Ban Facial Recognition Technologies for Children—and for Everyone Else*, 26 B.U. J. Sci. & Tech. L. 223, 277 (2020). *But see* Bruce Schneier, *We're Banning Facial Recognition. We're Missing the Point.*, N.Y. Times (Jan. 20, 2020), https://www.nytimes.com/2020/01/20/opinion/facial-recognition-ban-privacy.html (pointing out that facial recognition is only "one identification technology among many").

99. *See* Shervin Minaee et al., Biometric Recognition Using Deep Learning: A Survey (Feb. 8, 2021) (preprint), https://arxiv.org/pdf/1912.00271.pdf [https://perma.cc/3RL2-X3FE].

100. *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Calif. L. Rev. 1753, 1772–73 (2019) ("Thanks to deep-fake technology, an individual's face, voice, and body can be swapped into real pornography. . . . When victims discover that they have been used in deep-fake sex videos, the psychological damage can be profound—whether or not this was the video creator's aim."); Mary Anne Franks & Ari Ezra Waldman, *Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions*, 78 Md. L. Rev. 892, 893 (2019) ("Like other forms of nonconsensual pornography, digitally manipulated pornography turns individuals into objects of sexual entertainment against their will, causing intense distress, humiliation, and reputational injury."); *see also* Ashcroft v. Free Speech Coalition, 535 U.S. 234, 241–42 (2002) (distinguishing computer-generated images, which do not involve or harm any underlying person, from computer-morphed images, which do implicate the interests of real persons).

101. *See* Pasquale, *supra* note 65, at 1926; Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. Rev. 93 (2014).

102. *See* Moore v. Regents of Univ. of Calif., 793 P.2d 479 (Cal. 1990).

103. *See* Matthew Sag, *Copyright Safety for Generative AI*, 61 Hous. L. Rev. (forthcoming 2024); Benjamin L. W. Sobel, *Artificial Intelligence's Fair Use Crisis*, 41 Colum. J.L. & Arts 45 (2017). *But see* Mark A. Lemley & Bryan Casey, *Fair Learning*, 99 Tex. L. Rev. 743 (2021).

104. *See* Helen Nissenbaum, Privacy in Context (2009).

105. *See* Martin Abadi et al., *Deep Learning with Differential Privacy*, 2016 ACM Conf. on Comput. & Commc'ns Sec. 308.

106. *See* Sandra G. Mayson, *Bias in, Bias Out*, 128 Yale L.J. 2218, 2238 (2019) ("[I]f the base rate of the predicted outcome differs across racial groups, it is impossible to achieve (1) predictive parity; (2) parity in false-positive rates; and (3) parity in false-negative rates at the same time . . . . Race neutrality is not attainable."); Anupam Chander, *The Racist Algorithm?*, 115 Mich. L. Rev. 1023, 1036–37 (2017); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 680 (2016) ("As computer science scholars explain, biased training data leads to

is predictive policing, where AI techniques are being used to formulate criminal risk assessments both at the community level and at the individual level.[107] Many commentators have observed that policing practices and criminal law enforcement have been—and continue to be—pervaded by disparate treatment of protected classes, including race and gender.[108] Accordingly, they argue, any data drawn from past policing practices will necessarily reflect those same biases, and should be disqualified from informing future policing practices.[109] Other commentators have extended that critique to myriad contexts, including employment,[110] healthcare,[111] and consumer credit,[112] where past practices have been problematic. In these socially fraught areas, the criticism

discriminatory models."); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1104 (2019) ("[A] racial equity analysis of algorithmic criminal justice should not be a comparative one. . . . The mere fact that the status quo ante is characterized by racial injustice does not legitimatize proposals that preserve or extend some substantial part of that injustice."); Emily M. Bender et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, 2021 PROC. ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 610, 614 ("In accepting large amounts of web text as 'representative' of 'all' of humanity we risk perpetuating dominant viewpoints, increasing power imbalances, and further reifying inequality.").

107. *See* Andrew Guthrie Ferguson, *Policing Predictive Policing*, 94 WASH. U. L. REV. 1109 (2017); Elizabeth E. Joh, *Policing By Numbers: Big Data and the Fourth Amendment* 89 WASH. L. REV. 35 (2014); Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947; Megan Stevenson, *Assessing Risk Assessment in Action*, 103 MINN. L. REV. 303 (2018); Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System*, 52 U.C. DAVIS L. REV. 1067 (2018).

108. *See* Huq, *supra* note 106, at 1105–11; Andrew Guthrie Ferguson, *Illuminating Black Data Policing*, 15 OHIO ST. J. CRIM. L. 503 (2018); Shima Baradaran, *Race, Prediction, and Discretion*, 81 GEO. WASH. L. REV. 157 (2013); Jessica M. Eaglin, *Constructing Recidivism Risk*, 67 EMORY L.J. 59 (2017); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014) (gender).

109. *See* BERNARD E. HARCOURT, AGAINST PREDICTION: PROFILING, POLICING, AND PUNISHING IN AN ACTUARIAL AGE (2007); Ferguson, *Policing Predictive Policing*, *supra* note 107, at 1149 (addressing the argument that, "[i]f the underlying data is biased, then how can a data-driven system based on that data not also be biased?"); Sean Allan Hill II, *Bail Reform and the (False) Racial Promise of Algorithmic Risk Assessment*, 68 UCLA L. REV. 910, 944–45 (2021) (arguing that criticisms of pretrial risk assessment instruments do not go far enough in "interrogat[ing] how criminal laws and practices sustain prevailing beliefs of Black criminality," and that this technology "lends legitimacy to dangerousness predictions and thus encourages continued investments in the criminal legal system"); *see also* Huq, *supra* note 106, at 1131 (adopting a cost-benefit approach but arguing that "the operation of criminal justice coercion generates asymmetrical harms to black families and black communities").

110. *See* Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Charles A. Sullivan, *Employing AI*, 63 VILL. L. REV. 395 (2018); Jon Kleinberg et al., *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 1 (2018).

111. *See* Robin C. Feldman et al., *Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know*, 30 STAN. L. & POL'Y REV. 399 (2019); W. Nicholson Price II, *Black-Box Medicine*, 28 HARV. J.L. & TECH. 419 (2015); Pasquale, *supra* note 65, at 1926.

112. *See* Kristin Johnson et al., *Artificial Intelligence, Machine Learning, and Bias in Finance: Toward Responsible Innovation*, 88 FORDHAM L. REV. 499 (2019); PAM DIXON & ROBERT GELLMAN, THE SCORING OF AMERICA: HOW SECRET CONSUMER SCORES THREATEN YOUR PRIVACY AND YOUR FUTURE (2014).

is that the data at issue is fundamentally illegitimate for use in AI applications. Rehabilitation through procedural mechanisms is not a suitable option.[113]

C. Testing

In an important way, testing of AI models is far more orderly than testing of conventional software systems. Because the AI model is—or should be—a mathematical approximation of a naturally occurring pattern, the primary inquiry is whether the approximation is a close enough fit to the real phenomenon. A range of testing metrics allows AI modelers to estimate and compare an AI model's relative performance.

For classification problems, the simplest and most popular metric involves use of available data to determine the AI model's "accuracy." The full dataset is partitioned into a training and test set.[114] This partition allows the AI modeler to evaluate how well the AI model is likely to perform on new, previously unseen inputs.[115] A general rule of thumb is to allocate fifty to eighty percent of the original dataset to the training set, and leave the remainder for the test set. As long as the test set is independently drawn and adequately representative of the real world,[116] running the test set through the AI model offers simple heuristics that

113. *Cf.* Robin West, *The Limits of Process*, *in* GETTING TO THE RULE OF LAW 40–41 (2011) ("[W]e should acknowledge, before championing too loudly the cause of proceduralism, that excessively precious procedures in the face of grotesque substantive law from which there is truly no exit, even with all the procedure in the world, can be a massive insult to dignity. . . . In Hell, as Grant Gilmore observed, there will be perfect procedural justice."). *But see* Strandburg, *supra* note 91, at 1881 ("[P]reemptively depriving society of all such tools for all purposes in all significant decision contexts seems questionable as a policy matter, given the advantages of machine-learning-based decision tools in some contexts.").

114. *See* Lehr & Ohm, *supra* note 18, at 685 ("One key method is to randomly split, or partition, an entire dataset into two: a 'training' dataset and a 'test' dataset. A machine-learning algorithm is trained and learns the optimal predictive rules on the former. Then, the algorithm's accuracy and other performance metrics are assessed by asking it to predict the outcomes of the subjects in the latter. In this way, an algorithm is forced to predict data it has not 'seen' before . . . ."). The data partition can also include a validation set, which is used to tune the hyperparameters of the learning algorithm.

115. *See* GOODFELLOW ET AL., *supra* note 39, at 107 ("The central challenge in machine learning is that our algorithm must perform well on *new, previously unseen* inputs—not just those on which our model was trained. The ability to perform well on previously unobserved inputs is called generalization.").

116. Additionally, the training set and test set must have identical distributions, and they must be kept independent. *See* GOODFELLOW ET AL., *supra* note 39, at 108 ("We typically make a set of assumptions known collectively as the i.i.d. assumptions. These assumptions are that the examples in each dataset are independent from each other, and that the training set and test set are identically distributed, drawn from the same probability distribution as each other. . . . [These] assumptions enable[] us to mathematically study the relationship between training error and test error."). *But see* Sayash Kapoor & Arvind Narayanan, *Leakage and the Reproducibility Crisis in Machine-Learning-Based Science*, 4 PATTERNS, no. 9, Sept. 2023, at 1, 2, https://www.cell.com/

simulate how well the AI model is likely to perform on future data.[117] To be sure, an important limitation of this heuristic is that it is based only on known data, so maximizing the accuracy metric can yield worse performance due to overfitting of insufficient data. A common workaround is to "early stop" the training process before the accuracy metric reaches one-hundred percent.

Other available testing metrics offer alternative computations of best fit. Classification metrics such as "precision" and "recall" are optimized for detection of false positives or true positives, which may be more meaningful than overall accuracy in certain contexts.[118] For regression tasks involving continuous variables—rather than true-false classifications—a distance metric such as "mean squared error" offers a more appropriate computation of how close the AI model's estimated mapping is to the true natural pattern. Many more statistical methods offer a well-established canon of techniques to quantitatively compare the relative performance of AI models on a given task.[119]

These testing metrics are effective because the AI model has inherent smoothness properties. The data-driven, bottom-up training process relies on an implicit assumption that the data has a well-ordered, discoverable pattern. By definition, deep neural networks are mathematical representations of smooth, continuous functions capable of representing any arbitrary pattern.[120] That continuous property means selective sampling can offer useful evaluation insights.

By contrast, software is discontinuous; testing of conventional software systems tends to be haphazard, and generally lacks formal metrics.[121] Software is based on artificial, human-made constructs, rather

---

patterns/pdf/S2666-3899(23)00159-9.pdf (documenting data leakage as "a leading cause of errors in ML applications").

117. *See* Jie M. Zhang et al., *Machine Learning Testing: Survey, Landscapes and Horizons*, 48 IEEE Transactions on Software Eng'g 1, 6 (2022).

118. *See* Brendan Juba & Hai S. Le, *Precision-Recall versus Accuracy and the Role of Large Data Sets*, 33 AAAI Conf. on Artificial Intelligence 4039 (2019).

119. *See* Ben Hutchinson et al., *Evaluation Gaps in Machine Learning Practice*, 2022 ACM Fairness Accountability Transparency 1859, 1863, 1873 app. A (listing twenty-six commonly cited metrics across nine categories, including accuracy, precision, recall, F-score, overlap, distance, and AUC (area under the curve)).

120. *See* Goodfellow et al., *supra* note 39, at 192–93 (describing the universal approximation theorem, which states that a deep neural network can represent any kind of regular pattern as a continuous Borel measurable function).

121. *See, e.g.*, Steven B. Lipner, *The Birth and Death of the Orange Book*, IEEE Annals of the Hist. of Computing, Apr.–June 2015, at 19, 29 (noting the discovery that software projects evaluated for computer security per the Orange Book standard "fared no better under attack than any others"); Choi, *supra* note 22, at 583; Antonia Bertolino, *Software Testing Research: Achievements, Challenges, Dreams*, 2007 Future of Software Eng'g 85, 92 (citing a survey study that found "over half of the existing (testing technique) knowledge is based on impressions and perceptions and, therefore, devoid of any formal foundation" (citing Natalia Juristo et al., *Reviewing 25 Years*

than natural patterns like deep AI. Moreover, because of software's abstraction away from physical materials, software is unlike structural engineering or industrial manufacturing in that it lacks smoothly interpolated attributes such as compressive strength that can be adequately tested by selective sampling.[122] Instead, software failures can occur in completely arbitrary fashion, at any point along the control path. That lack of interpolability means the only way to ensure a program has no critical errors is to test every single control path that the code allows, as well as every functionality that the design requires. Yet, the exponential nature of software program complexity makes that task computationally impossible.[123]

Software testers have looked for heuristics that can reduce the amount of testing needed, with little success. For example, "code coverage" techniques seek to generate enough test cases to cover one-hundred percent of the source code—which sounds impressive but lacks any rigor as to how mere coverage equates with correctness.[124] Other methods seek to reduce the number of test cases by focusing attention on the most commonly known types of faults, through techniques such as equivalence partitioning, boundary value analysis, or all-pairs testing.[125] These latter techniques can be effective at detecting "known knowns," but are not aimed at locating unknown errors. Some scholars have championed model-based testing, which seeks to streamline testing by requiring the software code to conform to a formal model of the system's functionality.[126] Although this approach offers some theoretical potential, it has had minimal uptake in real-world practice due to limitations in formal

---

of Testing Technique Experiments, 9 Empirical Software Eng'g 7 (2004))); Ina Schieferdecker & Andreas Hoffmann, *Model-Based Testing*, *in* Encyclopedia of Software Engineering 556, 561 (Phillip A. Laplante ed., 2011) (observing that the "lack of quality metrics leads most companies to simply count the number of defects that emerge when testing occurs," and that "[f]ew organizations engage in other advanced testing techniques").

122. *See* Charles C. Mann, *Why Software Is So Bad*, MIT Tech. Rev., July/Aug., 2002, at 33, 36.

123. *See* Bertolino, *supra* note 121, at 91 (noting that seminal work in software testing theory provides "logical arguments to corroborate the quite obvious fact that testing can never be exact" and, furthermore, that there is "little guidance about what it is then that we can conclude about the tested software after having applied a selected technique" (citing Edsger W. Dijkstra, Notes on Structured Programming (2d ed. 1970))).

124. *See* Laura Inozemtseva & Reid Holmes, *Coverage Is Not Strongly Correlated with Test Suite Effectiveness*, 36 Int'l Conf. on Software Eng'g 435 (2014) (finding that code coverage is not a good proxy of test suite effectiveness); Hadi Hemmati, *How Effective Are Code Coverage Criteria?*, 2015 IEEE Int'l Conf. on Software Quality, Reliability & Sec. 151.

125. *See, e.g.*, Schieferdecker & Hoffmann, *supra* note 121, at 559; D. Richard Kuhn et al., *Software Fault Interactions and Implications for Software Testing*, 30 IEEE Transactions on Software Eng'g 418 (2004).

126. *See* Schieferdecker & Hoffmann, *supra* note 121, at 556.

modeling methods.[127] Ultimately, the lack of interpolability and the lack of metrics means that software testers typically rely on ad hoc practices, based on intuition and trial-and-error, to determine how to search for software bugs in the proverbial haystack.[128]

For now, testing methods for AI models remain imperfect and continue to face important limitations. The existing metrics have known shortcomings that often go unexamined.[129] In particular, most standard metrics are based on historical data that may not be adequately representative of the real world. It is far more challenging and costly to develop metrics that validate the novel, generative outputs of an AI system.[130] Robustness remains a substantial problem, as real-world performance is often worse than the testing metrics would predict.[131] Adversarial attacks add an extra factor: AI models can exhibit strange behaviors when probed in malicious ways.[132] To compensate, entities often resort to "online training," whereby the AI model is updated in real time with new data, but this adaptive approach raises difficult questions of how to validate ongoing changes to the model.[133] More

127. *See* Bertolino, *supra* note 121, at 93 (noting that "industrial adoption of model-based testing remains low and signals of the research-anticipated breakthrough are weak"); Schieferdecker & Hoffmann, *supra* note 121, at 556, 568 (stating that model-based testing is "rarely used in industrial-grade processes" and that adoption is "slow").

128. *See* Dudekula Mohammed Rafi et al., *Benefits and Limitations of Automated Software Testing: Systematic Literature Review and Practitioner Survey*, 7 Int'l Workshop on Automation of Software Test 36 (2012) (finding that automated software tools are unlikely to fully replace manual testing).

129. *See* Nathalie Japkowicz, *Why Question Machine Learning Evaluation Methods?*, AAAI Workshop on Learning from Imbalanced Datasets (2006).

130. This issue is called the test oracle problem. *See* Zhang et al., *supra* note 117, at 8 ("Currently, the identification of test oracles remains challenging, because many desired properties are difficult to formally specify. Even for a concrete domain specific problem, the oracle identification is still time-consuming and labour-intensive, because domain-specific knowledge is often required.").

131. *See* Shibani Santurkar, Machine Learning Beyond Accuracy: A Features Perspective On Model Generalization (Sept. 2021) (Ph.D. dissertation, MIT), https://dspace.mit.edu/bitstream/handle/1721.1/139920/Santurkar-shibani-PhD-EECS-2021-thesis.pdf (explaining that benchmark performance turns out to be remarkably brittle in real-world performance, a problem illuminated by adversarial examples); Evan Elwell, Nat'l Acads. of Scis., Eng'g & Med., Testing, Evaluating, and Assessing Artificial Intelligence-Enabled Systems Under Operational Conditions for the Department of the Air Force: Proceedings of a Workshop—in Brief 1, 7 (2023) (noting that often "performance of the deployed system in the operational domain was much worse than predicted during the test phase").

132. *See* Apostal Vassilev et al., NIST, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations 54 (2024) ("Unfortunately, it is not possible to simultaneously maximize the performance of the AI system with respect to these attributes. For instance, AI systems optimized for accuracy alone tend to underperform in terms of adversarial robustness and fairness. Conversely, an AI system optimized for adversarial robustness may exhibit lower accuracy and deteriorated fairness outcomes."); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. Rev. 1315, 1351 (2020) (discussing examples).

133. *See, e.g.*, FDA, Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) (2021).

broadly, these quantitative metrics may fail to detect problems of apophenia, where the AI modeler attempts to represent an illusory pattern.[134] Additionally, many commentators have articulated overarching concerns that AI testing may not capture qualitative values that cannot be easily quantified.[135]

In short, AI testing is not yet a mature science and much more work is needed to develop improved metrics. Nonetheless, if the point of comparison is software testing, then the differential factor is that AI testing offers more potential for objective evidence regarding the degree of care exercised by the AI modeler.[136]

III. Bad AI Outcomes

The second factor that is relevant to the application of professional malpractice is whether there are serious harms that are statistically unavoidable because of the lack of scientific precision or control. While some AI modeler errors may be obvious or common knowledge, plenty of others are likely to be unintuitive and surprising.[137] The key takeaway here is that harmful outcomes are an expected feature even of competent AI modeling work.

Harmful outcomes can be mapped along two major axes. One axis distinguishes malintent versus accidental harm. Many AI systems exhibit unintended behaviors that could result in injury. By contrast, some AI systems may be released with knowledge or intent to cause some legal harm. An extension of such cases involves foreseeable misuse by end users, which could be attributed back to the AI vendor in the form of

134. *See* Ifeoma Ajunwa, *Automated Video Interviewing as the New Phrenology*, 36 Berkeley Tech. L.J. 1173, 1187 (2021) (analogizing the use of AI to predict emotions or character based on facial features to the fake science of phrenology); Lehr & Ohm, *supra* note 18, at 674–75 (explaining that "the decisionmaker must translate the predictive goal to a specified outcome variable," and that it is too easy to "lose sight of the intrinsic limit of targeting policy only on what we can measure").

135. *See* Selbst, *supra* note 132, at 1338 ("[W]here the entire purpose of an AI system is to predict the unobservable, there may be no way to know how far off the approximation is."); danah boyd & Kate Crawford, *Critical Questions for Big Data*, 15 Info. Comm. & Soc'y 662, 667 (2012); Kaminski, *supra* note 2, at 1397–98; *see also* Cary Coglianese, *The Limits of Performance-Based Regulation*, 50 U. Mich. J.L. Reform 525, 562 (2017) ("If the performance required is unrelated to the desired outcomes, or if it is too broadly defined so that firms can comply in ways that will have no impact on the desired outcome, then these standards will fail. In addition, problems that performance standards seek to address may still persist if other factors, unaffected by the regulation, contribute to the problem or if the regulation fails to address the root causes of the problem.").

136. *See* Notice of Artificial Intelligence Safety Institute Consortium, 88 Fed. Reg. 75276 (Nov. 2, 2023) (noting that the consortium will be responsible for, inter alia, developing new benchmarks, testing environments, and red-teaming methods).

137. *See* Selbst, *supra* note 132, at 1342–46; Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. Chi. L. Rev. 1311, 1313 (2019); Ryan Calo, *Is the Law Ready for Driverless Cars?*, Commc'n ACM, May 2018, at 34, 35.

either duty of care or constructive knowledge. The second axis distinguishes between harms caused by incorrectness or falsity and harms caused by perpetuation of bona fide patterns. At one end, harm arises because the AI system's "objective function" is misaligned with the AI user's true goals. At the opposite end, harm is alleged not because of incorrect modeling or use, but because the AI system accurately reflects or accelerates existing patterns of societal harm. Some quick caveats: this taxonomy is not intended to delineate rigid categories of harm, but instead to facilitate more precise diagnoses of different etiologies of AI harm. In fact, any given AI system might trigger multiple notions of harm at once. Moreover, this taxonomy focuses only on outputs of AI systems, rather than on harms arising purely at the input stages.

TABLE 1

	Incorrectness	**Perpetuation**
Accident	Self-driving collisions	Recommendation algorithms
Intent	Deepfake porn	Discriminatory profiling
Foreseeable Misuse	Adversarial attacks	Plagiarism

The computer science literature has focused most heavily on the upper-left quadrant, where the AI modeler intends no harm, and harms that do occur are attributable to failures of accuracy or alignment.[138] For example, when an autonomous vehicle crashes, it is typically because there is a failure in object detection, path planning, or avoidance of other road users' errors.[139] In most conventional cases, the automaker has not programmed its vehicle with deliberate intent to cause collision.[140] Likewise, if a facial recognition algorithm identifies the wrong individual, or fails to identify individuals of certain ethnicities, the harm is likely due to inadvertent inaccuracy rather than purposeful misdirection.

Incorrectness is endemic to AI methodologies. Because the AI model is, at best, an approximation of the real world, it is inevitable that

138. *See* Dario Amodei et al., Concrete Problems in AI Safety (July 25, 2016) (preprint), https://arxiv.org/pdf/1606.06565.pdf [https://perma.cc/9GNL-BDL7].

139. *See* Matthew Wansley, *The End of Accidents*, 55 U.C. DAVIS L. REV. 269, 281 (2021) (explaining that autonomous vehicle software includes mapping, behavior prediction, and motion planning functions).

140. *But see* Samuel Judson et al., *'Put the Car on the Stand': SMT-based Oracles for Investigating Decisions*, 2024 ACM SYMPOSIUM ON COMPUT. SCI. & L. (forthcoming 2024), https://arxiv.org/pdf/2305.05731.pdf [https://perma.cc/C5SP-K6NR] (offering a formal method of evaluating an automated decision maker's "intent," based on its functional behavior, to distinguish between "normal," "impatient," and "pathological" vehicles).

mismatches will arise.[141] Many of those failures are attributable to avoidable implementation errors, but at least some are due to initialization choices that—as explained in the prior section—have no robust justification other than past practice and guesswork. Moreover, an AI model is a snapshot, so it can become outdated if there is change over time, such as a natural language model that is unaware of recent news events or semantic drifts,[142] or change in context, such as a medical AI system trained in a high-resource hospital setting and transferred to a low-resource setting.[143]

The latent risk of AI incorrectness can convert to intentional harm when there is intent to design or use an AI system in a false manner. Such misuse might result in charges of fraud or misrepresentation, as when the AI modeler knowingly overstates the capabilities of the AI model.[144] So-called "hallucinations" of large language models might belong in this category as well.[145] More insidious are cases such as apps for deepfake porn where the design of the AI system is intended to create content that is deceptive. Used maliciously, "deepfake" text, images, and videos can cause reputational harm to individual persons.[146] At scale, AI-generated fake content can promote distrust in journalism, social institutions, and truth itself.[147] These threats are not new, but AI tools are likely to make the impact much greater.[148]

---

141. Inioluwa Deborah Raji et al., *The Fallacy of AI Functionality*, 2022 ACM Conf. on Fairness Accountability & Transparency 959, 962 (describing four ways that AI systems can fail to function: impossible tasks, engineering failures, post-deployment failures, and communication failures); Strandburg, *supra* note 91, at 1861 (explaining the "trade-off between using an outcome variable for which 'bigger' data is available and using a better proxy for the true criteria of interest").

142. *See, e.g.*, *What is ChatGPT?*, OpenAI, https://help.openai.com/en/articles/6783457-what-is-chatgpt [https://perma.cc/SKB8-2SCF] (noting that ChatGPT "has limited knowledge of world and events after 2021").

143. *See* W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 Harv. J.L. & Tech. 65, 68 (2019).

144. *See* Dave Michaels & Rebecca Elliott, *SEC, DOJ Probe Tesla Over Statements About Autopilot Functionality*, Wall St. J. (Oct. 27, 2022, 3:23 PM), https://www.wsj.com/articles/sec-doj-probe-tesla-over-statements-about-autopilot-functionality-11666898610.

145. *See* Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. Free Speech L. 489, 499 (2023).

146. *See* Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 Calif. L. Rev. 1753, 1774 (2019) ("[D]eep-fake technology can be used to harm victims along other dimensions due to their utility for reputational sabotage. Across every field of competition—workplace, romance, sports, marketplace, and politics—people will have the capacity to deal significant blows to the prospects of their rivals.").

147. *See* Paul Ohm, Ayelet Gordon-Tapiero & Ashwin Ramaswami, *Fact and Friction: Mandating Friction to Fight False News*, 57 U.C. Davis L. Rev. 171 (2023).

148. *Cf.* Mary Anne Franks, *Unwilling Avatars: Idealism and Discrimination in Cyberspace*, 20 Colum. J. Gender & L. 224, 255–56 (2011) (setting forth four distinctive features of cyberspace that exacerbate its harms: anonymity, amplification, permanence, and virtual captivity).

It is also useful to consider cases where harm is caused primarily by third-party uses, but the risk of harm is foreseeable enough that a duty of care can be attributed to the AI modeler.[149] The canonical case to fit this pattern would be an adversarial attack that exploits a hidden weakness in the AI model in order to generate incorrect behavior.[150] As with conventional cyber attacks, however, there will likely be substantial difficulties in proving that a particular exploit was both foreseeable and remediable prior to the attack.

An entirely different set of concerns involves those where incorrectness is not the root cause of injury. Instead, the identified harm is one already extant in society, and is reflected and amplified by the AI system. For example, recommendation systems are used to provide personalized nudges that steer individuals toward their prior preferences. As numerous scholars have brooded, too much of a good thing can lead to undesirable outcomes.[151] Credit scoring offers another helpful illustration: even assuming arguendo that there are no errors in the predictions of creditworthiness, a more fundamental set of objections is that such scores reinforce existing social inequities in undesirable ways. Replicating and streamlining existing patterns of social behavior can accelerate unintended negative effects at scale.[152]

When the historical data is known to be problematic, the use of AI methods to replicate that existing practice could constitute intent to discriminate. For example, an employer might use a hiring system that improperly favors men over women.[153] Arguably, bad intent could extend to the AI modeler who makes training choices that cause the AI model to align with one gender rather than with the actual job credentials.[154] In a similar vein, developers of automated policing systems have been criticized for bringing products to market that are biased against underserved communities. Here, too, the critique sounds not just in inaccuracy,[155] but also with a heavy suggestion that the bias is—if not intentional—willfully blind.[156]

---

149. *See In re* Sept. 11 Litig., 280 F. Supp. 2d. 279, 296, 313 (S.D.N.Y. 2003).

150. *See generally* Vassilev et al., *supra* note 132.

151. *See* Cass Sunstein, Republic.com 3 (2001); Joseph Turow, The Daily You (2011).

152. *See* Lemley & Casey, *supra* note 137, at 1339 (observing "widespread concerns" that AI systems could "create negative feedback loops that are hard to break"); Kolt, *supra* note 1 (manuscript at 46–49) (worrying about systemic risks that cause the "gradual erosion of social and political institutions and values").

153. *See* Pauline T. Kim & Matthew T. Bodie, *Artificial Intelligence and the Challenges of Workplace Discrimination and Privacy*, 35 ABA J. Labor & Employment L. 289, 294 (2021).

154. *See* Pauline T. Kim, *Manipulating Opportunity*, 106 Va. L. Rev. 867 (2020).

155. *See* Ferguson, *supra* note 108, at 514 ("If the crime data collected from particular areas becomes the only data in the system, then police data systems will mirror police patrols, not necessarily actual crime rates.").

156. *Id.* at 516 ("This is not to say that predictive policing is intentionally racially discriminatory, but only that, like traditional policing, it suffers from implicit and explicit racial biases, and tracks the structural problems inherent in policing.").

The final box extends notions of AI harm to scenarios where third-party misuse of the AI model causes harm by hewing too closely to the modeled pattern. A leading example is the use of generative AI systems to create plagiarized content. Several lawsuits have alleged that large language models reproduce texts or images that are strikingly similar to copyrighted content used as training data on an unlicensed basis. Moreover, even when the system is not overtly copying other authors' content, the use of such systems may breach obligations to produce original work in a broad range of contexts such as classroom assignments, journalism, and book publishing. A different set of examples involves third-party uses that are individually correct but that generate bad systemic interactions. For example, algorithmic financial trading has been known to generate market distortions such as flash crashes.[157]

In sum, AI experts expect that their work will produce a broad range of harmful effects. Some of these outcomes may be avoidable—especially those that fall within intentional harms—but many others may not be. An important area of further inquiry will be to distinguish the types of AI harms that are unforeseeable from those that are reasonably foreseeable—or perhaps even obvious or commonly known.[158]

IV. Essential Services

Thus far, I have argued that the case for treating AI modelers as a profession is a closer one than it is for software developers, albeit one that still leans in favor for now. Much of the work involved in training neural networks is either menial or guided by well-established mathematical principles. Nevertheless, an important component of the work involves subjective judgments that are guided by customary practices derived from trial-and-error, rather than an objective understanding of why those choices work well. Moreover, the risks of harm from AI systems are quite significant.

The third factor further complicates that assessment: do AI modelers perform an essential societal service even when their customary practices cause harm? Leading voices within the AI community—including top luminaries of the field—have cautioned that AI deployment should be slowed down, because the dangers posed by modern AI methods could pose an existential threat to human society.[159] Other prominent

157. *See* Gina-Gail S. Fletcher, *Deterring Algorithmic Manipulation*, 74 Vand. L. Rev. 259, 262–63 (2021).

158. *See* Selbst, *supra* note 132, at 1342 ("Much of the existing research points to foreseeability as the greatest challenge that AI poses for tort law.").

159. *See Statement on AI Risk*, Ctr. for AI Safety, https://www.safe.ai/statement-on-ai-risk#open-letter [https://perma.cc/N856-E3J6] (The one-sentence statement reads: "Mitigating the

voices have criticized the cost-benefit tradeoff, citing environmental costs, labor disruptions, and data privacy harms.[160] Polls suggest public support for AI is weakening.[161] To be sure, this view is not unanimous, but it raises the question whether an occupation should be trusted with self-governance, when many of its most prominent leaders believe that self-governance cannot work.[162]

Another salient feature of the AI modeler community is that its membership remains small and exclusive, and it has not yet exploded in size the way the software developer community has.[163] In part, the deep learning revolution is still young, and barriers to entry have been high—only a handful of entities in the world have had the compute power to build state-of-the-art neural networks.[164] Some commentary has argued that new techniques are rapidly reducing barriers to entry,[165] but it remains to be seen how those predictions will bear out. For now, the community remains exclusive. Smallness cuts both ways. On the one hand, it is easier to draw boundaries around the occupation and determine who is in or out. Consensus on best practices may be easier to build. On the other hand, it casts doubt on the urgency and criticality of the work, because it suggests there is not yet a broad societal dependency on AI services.

Unlike medical or legal services—or even software services—whose absence causes acute hardships, deep learning AI models are only

---

risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."); *see also* Pause Giant AI Experiments: An Open Letter, Future of Life Inst. (Mar. 22, 2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/ [https://perma.cc/W69V-RLJJ] (citing risks of disinformation, worker displacement, and "loss of control of our civilization" as reasons to pause research on state-of-the-art AI systems).

160. *See* Bender et al., *supra* note 106; Timnit Gebru & Margaret Mitchell, *We Warned Google that People Might Believe AI Was Sentient. Now It's Happening.*, Wash. Post (June 17, 2022), https://www.washingtonpost.com/opinions/2022/06/17/google-ai-ethics-sentient-lemoine-warning/.

161. *See* Alec Tyson & Emma Kikuchi, *Growing Public Concern About the Role of Artificial Intelligence in Daily Life*, Pew Rsch. Ctr. (Aug. 28, 2023), https://www.pewresearch.org/short-reads/2023/08/28/growing-public-concern-about-the-role-of-artificial-intelligence-in-daily-life/ [https://perma.cc/YN29-QVCB] (finding fifty-two percent of Americans are more concerned than excited about the increased use of AI, and only ten percent are more excited than concerned).

162. This scenario resembles the liar's paradox. *See* Douglas R. Hofstadter, Gödel, Escher, Bach 17 (1979) ("Epimenides was a Cretan who made one immortal statement: 'All Cretans are liars.'").

163. *See* David Kelnar, MMC Ventures, The State of AI 2019: Divergence 87 (2019), https://mmc.vc/resources/fund-brochures/The-MMC-State-of-AI-2019-Report.pdf [https://perma.cc/237P-S9LP] (estimating, in 2019, the global pool of AI talent to be "as few as 22,000 highly-trained AI specialists" or "up to 300,000 AI researchers and practitioners within broader technical teams").

164. *See* Gerrit De Vynck, *How Big Tech Is Co-opting the Rising Stars of Artificial Intelligence*, Wash. Post, (Sept. 30, 2023, 8:00AM), https://www.washingtonpost.com/technology/2023/09/30/anthropic-amazon-artificial-intelligence/.

165. *See* Dylan Patel & Afzal Ahmad, *Google "We Have No Moat, And Neither Does OpenAI"*, SemiAnalysis (May 4, 2023), https://www.semianalysis.com/p/google-we-have-no-moat-and-neither [https://perma.cc/YEY8-YEVA].

beginning to be ushered into common use. Many AI-based services are deployed as an enhancement of existing services, and their presence is otherwise unremarkable to ordinary citizens. When AI systems do attract attention, communal bans have become a familiar response.

For the time being, perhaps the most apt analogy is to air travel services during the early twentieth century, when aviation was not yet considered "common" or "essential."[166] Even though the potential advantages of air travel were obvious, the immaturity of avionics science led lawmakers to label air travel an ultrahazardous activity and to impose strict liability.[167] Likewise, until AI-based services prove to be indispensable, it is not obvious that AI experts should be given special deference as a matter of law.

New technologies have a habit of becoming old hat, however, and AI is no exception. Whenever AI becomes an ordinary fixture of everyday society, the basic inquiry will reemerge as to whether ordinary reasonable care is a viable standard or whether some alternative framework is needed. Central to that inquiry will be the two other factors discussed above: the degree of scientific uncertainty in the field, and the extent of bad outcomes ascribed to practitioners in good standing.

Conclusion

In prior work, I explained that courts have used the "professional" label to impose an alternate liability framework when courts have hesitated to trust jury sentiments, yet needed some form of judicial oversight for unfit practitioners. The professional malpractice doctrine gives deference to legitimate exercises of professional judgment, on the basis that the professional's work cannot be reduced to an exact science. I argued further that software developers should be treated as professionals—like medical care or legal practice, much of software development remains an inexact art.

Here, I have extended those earlier writings by explaining how modern AI work differs in key aspects from software development work, even though both ply code and data. In particular, AI modelers are engaged in automating the mathematical representation of a naturally occurring data pattern. The aim is to avoid injecting undue extrinsic interference into that learning process. That bottom-up synthesis stands

166. *See* Restatement of Torts § 520 cmt. g (Am. L. Inst. 1938).

167. *See* Henry Grady Gatlin, Jr., Note, *Tort Liability in Aircraft Accidents*, 4 Vand. L. Rev. 857, 861, 874 (1951) ("Aviation in its infancy became branded a highly questionable and dangerous enterprise. . . . With the increased technological development of aviation and an establishment of aviation as a safe mode of travel, the early doctrine of strict liability against the air carrier is disappearing.").

in sharp contrast with the top-down engineering that occurs in mainstream software development, where the end goal is to specify, design, and execute an arbitrarily devised human construct.

Thus, the case for AI professionals is a closer question. It is evident that AI work involves less code complexity and that there are fewer human decision points that need to be examined to assess fault. Yet, those decisions that do need to be made remain undertheorized, and are guided more by folk wisdom than by scientific understanding. As the knowledge of the field advances, it may become easier to apply an ordinary reasonable care framework to the work that AI modelers perform. But there is still a strong case today that deep learning theory has not matured enough to support an objectively reasonable standard of care.

Meanwhile, a more immediate policy question for courts is whether they believe deep learning AI practitioners offer services to society that are vital—or that are merely profitable. Ordinarily, the entity that creates a risk of harm is expected to bear responsibility for the injuries caused by its operations. If AI-based services are not essential to societal health (or are even detrimental to it), then it makes little sense to carve an exception to the default rule. After all, early courts understood well the transformative promise of air travel, yet still chose to require aircraft manufacturers and operators to bear full liability for all injuries caused by crashes.

Whether AI modelers are professionals or nonprofessionals will affect in turn the scope of ethical duties they must assume. On the one hand, if AI modelers are nonprofessionals, then AI ethics principles lack an effective enforcement mechanism unless they are converted into legal duties. In such a scenario, AI ethics principles would need to be translated into legal rules in order to have real purchase. On the other hand, if AI modelers are treated as professionals, then the law will enforce the customary practices among the AI community. It would increase greatly the impact of articulating consensus standards of professional ethics as an instrument to guide AI practitioners in their conduct and to guide courts in their liability decisions.