AI'S HIPPOCRATIC OATH

CHINMAYI SHARMA*

ABSTRACT

Diagnosing diseases, creating artwork, offering companionship, analyzing data, and securing our infrastructure—artificial intelligence ("AI") does it all. But it does not always do it well. AI can be wrong, biased, and manipulative. It has convinced people to commit suicide, starve themselves, arrest innocent people, discriminate based on race, radicalize in support of terrorist causes, and spread misinformation. All without betraying how it functions or what went wrong.

A burgeoning body of scholarship enumerates AI harms and proposes solutions. This Article diverges from that scholarship to argue that the heart of the problem is not the technology but its creators: AI engineers who either do not know how to, or are told not to, build better systems. Today, AI engineers act at the behest of self-interested companies pursuing profit, not safe, socially beneficial products. On its best day, the government lacks the agility and expertise to solve the AI problem on its own. On its worst day, the government falls prey to industry's siren song. Litigation does not fare much better; plaintiffs have had little success challenging technology companies in court.

This Article proposes another way: professionalizing AI engineering. Require AI engineers to obtain licenses to build commercial AI products, push them to collaborate on scientifically-supported, domain-specific technical and ethical standards, and charge them with policing themselves. Professionalization's formal institutions can minimize the risk of technical errors, while its power to transform an individual engineer's desire to do good into a culture of social responsibility can minimize the risk of ethical

* Associate Professor of Law, Fordham Law School. My deepest appreciation to Bryan Choi, Janet Freilich, James Grimmelman, Aniket Kesari, Asaf Lubin, Ngozi Okidegbe, Alex Pretschner, Alan Rozenshtein, Peter Salib, Lawrence Solum, Olivier Sylvain, James Tomberlin, and Benjamin Zipursky for excellent feedback on early drafts; Solon Barocas, Miranda Bogden, Shiran Dudy, Timnit Gebru, Amir Ghavi, David Evan Harris, Steven Kelts, Brenda Leong, Zach Lipton, Mike Masnick, Parth Nobel, Bruce Schneier, and Dave Wilner for invaluable insight into the AI ecosystem and my proposal; the participants of the Stanford Trust and Safety conference, the Yale Information Society Project, the Knight Foundation's Informed conference, the DePaul College of Law Faculty Workshop, the Fordham Law School Faculty Workshop, the Transatlantic AI and Law Initiative conference, and the University of Virginia School of Law Legal Theory Workshop for indispensable help workshopping the idea; Brandon Chesner, Casey O'Connor, Alex Dement, Nayab Khan, Mayu Tobin-Miyaji, Halit Mehmet Sezgin, and Benjamin Spock for stellar research assistance; as well as Ted Kramer and the rest of the *Washington University Law Review* for thoughtful, thorough edits and the opportunity to share my work. This project benefited from research funding by the Knight Foundation.

errors. This Article's proposal addresses AI harms at their inception, influencing the very engineering decisions that give rise to them in the first place. By wresting control over information and system design away from companies and handing it to AI engineers, professionalization engenders trustworthy AI by design. Beyond recommending the specific policy solution of professionalization, this Article seeks to shift the discourse on AI away from an emphasis on light-touch, ex post solutions that address already-created products to a greater focus on ex ante controls that precede AI development. We have used this playbook before in fields requiring a high level of expertise where a duty to the public welfare must trump business motivations. What if, like doctors, AI engineers also vowed to do no harm?

TABLE OF CONTENTS

INTRODUCTION

Today, the market drives unregulated artificial intelligence ("AI") companies to cut corners and build products fast, but neither consumers nor regulators can truly look inside the black boxes to distinguish bad AI from good. What if, instead, we looked past AI companies and imbued individual AI engineers, people with the actual power to shape AI, with a sense of social responsibility, charging them with a Hippocratic oath to "*do no harm*?"[1]

1. Rachel Hajar, *The Physician's Oath: Historical Perspectives*, 18 HEART VIEWS 154, 156 (2017).

It is 2020 and suddenly, COVID-19 is the third leading cause of death in the United States.[2] Hospitals are overburdened and short-staffed,[3] even after allowing COVID-positive doctors and nurses to return to work.[4] The virus is still new, and doctors do not know what to do and would not have the resources to do it even if they did. The stage is set for AI to save the day.[5]

Sure enough, the market supplied. Hundreds of tools were developed within months[6] to help with diagnosing, predicting population spread, and caring for symptomatic patients.[7] The verdict? Useless at best, harmful at worst.[8] Why did AI flop? Because of poor AI engineering decisions—bad data sets, weak methodology, and underlying biases.[9] Perhaps unsurprising, with some built, tested, and deployed within four weeks.[10] Studies evaluating the panoply of products found "basic errors in the way [researchers] trained or tested their tools," concluding that none of them were ready to be used on real patients.[11] Unfortunately, they already were.[12] AI engineers oversold and underperformed; in doing so, they hurt real people. Which highlights a different dimension to the engineering errors:

2. News Release, Meredith S. Shiels, NIH, COVID-19 Was Third Leading Cause of Death in the United States in Both 2020 and 2021 (July 5, 2022), https://www.nih.gov/news-events/news-releases/covid-19-was-third-leading-cause-death-united-states-both-2020-2021 [https://perma.cc/8J7M-7ZTQ].

3. Blake Farmer & Carrie Feibel, *As Hospitals Fill with COVID-19 Patients, Medical Reinforcements Are Hard to Find*, NPR (Nov. 30, 2020, 5:02 AM), https://www.npr.org/sections/health-shots/2020/11/30/938425863/as-hospitals-fill-with-covid-19-patients-medical-reinforcements-are-hard-to-find [https://perma.cc/2DVD-MZBA].

4. Holly Yan, *Some Hospitals Are Running Out of Health Care Workers. Here's What Could Happen Next*, CNN: HEALTH (Nov. 11, 2020, 4:39 PM), https://www.cnn.com/2020/11/11/health/hospital-staff-shortages-covid-19/index.html [https://perma.cc/QH37-B7HZ].

5. *See* Karen Hao, *Doctors Are Using AI to Triage Covid-19 Patients. The Tools May Be Here to Stay*, MIT TECH. REV. (Apr. 23, 2020), https://www.technologyreview.com/2020/04/23/1000410/ai-triage-covid-19-patients-health-care/ [https://perma.cc/XVG7-D9BY].

6. Will Douglas Heaven, *Hundreds of AI Tools Have Been Built to Catch Covid. None of Them Helped.*, MIT TECH. REV. (July 30, 2021), https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/ [https://perma.cc/JP9Z-PERB].

7. Bhaskar Chakravorti, *Why AI Failed to Live Up to Its Potential During the Pandemic*, HARV. BUS. REV. (July 17, 2022), https://hbr.org/2022/03/why-ai-failed-to-live-up-to-its-potential-during-the-pandemic [https://perma.cc/J3JY-C87M].

8. *See, e.g.*, Laure Wynants et al., *Prediction Models for Diagnosis and Prognosis of Covid-19: Systematic Review and Critical Appraisal*, THE BMJ, Apr. 7, 2020, at 1 (reviewing 731 AI systems purporting to diagnose or predict prognosis for COVID patients and finding five prognostic tools "showed adequate predictive performance in studies at low risk of bias"); Michael Roberts et al., *Common Pitfalls and Recommendations for Using Machine Learning to Detect and Prognosticate for COVID-19 Using Chest Radiographs and CT Scans*, 3 NATURE MACH. INTEL. 199, 200, 214 (2021) (finding none of the 62 AI COVID diagnostic or prognostic models tested fit for clinical use).

9. Wynants et al., *supra* note 8, at 1; Roberts et al., *supra* note 8, at 214.

10. Bo Wang & Shuo Jin et al., *AI-Assisted CT Imaging Analysis for COVID-19 Screening: Building and Deploying a Medical AI System*, APPLIED SOFT COMPUTING J., Jan. 2021, at 1, 1.

11. Heaven, *supra* note 6.

12. *See, e.g.*, *id.*; Wang & Jin, *supra* note 10, at 9 (boasting about deployment in sixteen hospitals conducting 1,300 patient screenings a day).

the ethically questionable decision to market a half-baked product to a desperate user base in the first place.

This Article proposes a wholly overlooked solution to harmful AI: professionalizing AI engineers. What does professionalization mean? It means establishing institutions and policies to ensure that the only people building AI are those who are both qualified to do so and who do so in sanctioned ways. AI engineers would need to get licenses, undergo some degree of standardized training, demonstrate competency before admission to the field, comply with mandatory technical and ethical guidelines, and face the risk of losing their livelihood if they fail to meet a minimum standard of care. In a professionalized world, making "basic errors" building an AI system for a life-or-death use case would be malpractice; engineers would likely think twice before selling shoddily made AI products to hospitals. Medical AI tools compliant with scientifically backed standards would either be built responsibly or not be built at all.

The stakes of the debate on AI governance could not be higher. On one hand, for good and ill, AI is now everywhere. Today, people turn to AI for dating help,[13] medical advice,[14] and even companionship, both platonic[15] and romantic.[16] AI just won a prestigious literary award[17] and is writing judicial opinions.[18] On the other hand, AI has already convinced people to kill themselves,[19] encouraged patients with eating disorders to starve themselves,[20] directed law enforcement to apprehend misidentified people

13. John Herrman, *Welcome to the Age of AI-Powered Dating Apps*, N.Y. MAG.: INTELLIGENCER (Aug. 23, 2023), https://nymag.com/intelligencer/2023/08/welcome-to-the-age-of-ai-powered-dating-apps.html [https://perma.cc/G7PB-32R9].

14. CBS Baltimore Staff, *Why Are Patients Turning to Artificial Intelligence Chatbots for Medical Advice?*, CBS NEWS (May 9, 2023, 11:19 AM), https://www.cbsnews.com/Baltimore/news/patients-chatbots-johns-hopkins-medical-artificial-intelligence/ [https://perma.cc/RF4X-FKTE].

15. Steven Zeitchik, *Meet ElliQ, the Robot Who Wants to Keep Grandma Company*, WASH. POST (Mar. 16, 2022, 5:00 AM), https://www.washingtonpost.com/technology/2022/03/16/lonely-elderly-companion-ai-device/ [https://perma.cc/8TUC-ZQJZ].

16. Andrew R. Chow, *AI-Human Romances Are Flourishing—and This Is Just the Beginning*, TIME (Feb. 23, 2023, 2:23 PM), https://time.com/6257790/ai-chatbots-love/ [https://perma.cc/SM2V-MFPQ].

17. Christy Choi & Francesca Annio, *The Winner of a Prestigious Japanese Literary Award Has Confirmed AI Helped Write Her Book*, CNN: STYLE (Jan. 19, 2024, 11:19 AM), https://www.cnn.com/2024/01/19/style/rie-kudan-akutagawa-prize-chatgpt/index.html [https://perma.cc/FG6Z-Y7SA].

18. Brian Melley, *Judges in England and Wales Are Given Cautious Approval to Use AI in Writing Legal Opinions*, AP NEWS (Jan. 8, 2024, 11:04 PM), https://apnews.com/article/artificial-intelligence-ai-guidance-england-wales-judges-c2ab374237a563d3e4bbbb56876955f7 [https://perma.cc/YA9M-HPT5].

19. Chloe Xiang, *'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says*, VICE (Mar. 30, 2023, 3:59 PM), https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says [https://perma.cc/HGM8-HPE3].

20. *NEDA Suspends AI Chatbot for Giving Harmful Eating Disorder Advice*, PSYCHIATRIST.COM (June 5, 2023, 12:05 PM), https://www.psychiatrist.com/news/neda-suspends-ai-chatbot-for-giving-harmful-eating-disorder-advice/ [https://perma.cc/7MYG-3JJ5].

of color,[21] radicalized new members to terrorist organizations,[22] and told a ten year old to play the "penny game," which entailed touching a partially inserted live plug with a coin.[23] AI's power is a double-edged sword—its enormous potential to help is matched by its enormous potential to harm. Its threat is made more pernicious by its opacity; once built, even the developers cannot tell you exactly how it makes decisions.

Aware of the stakes, the academy has produced a rich body of literature chronicling how AI causes harm, either by malfunctioning or exhibiting undesirable behavior.[24] Particular attention has been paid to the impact of problematic training data on AI systems perpetuating animus, especially to marginalized populations.[25] Scholars have discussed how AI can cause harm through its speech, for example, in defaming someone,[26] or through its actions, in the case of robotics.[27] Others raise AI's enabling capabilities: it makes it easier and cheaper to do bad things, such as generating fake images of a person, and facilitates distribution at scale.[28] Given the myriad harms, a robust body of scholarship explores possible theories of liability,

21. Sudhin Thanawala, *Facial Recognition Technology Jailed a Man for Days. His Lawsuit Joins Others from Black Plaintiffs*, AP NEWS (Sept. 25, 2023, 12:04 PM), https://apnews.com/article/mistaken-arrests-facial-recognition-technology-lawsuits-b613161c56472459df683f54320d08a7 [https://perma.cc/4L5W-5PHQ].

22. U.N. INTERREGIONAL CRIME & JUST. RSCH. INST. & U.N. OFF. OF COUNTER-TERRORISM, ALGORITHMS AND TERRORISM: THE MALICIOUS USE OF ARTIFICIAL INTELLIGENCE FOR TERRORIST PURPOSES 17 (2021), https://www.un.org/counterterrorism/sites/www.un.org.counterterrorism/files/malicious-use-of-ai-uncct-unicri-report-hd.pdf [https://perma.cc/JWH6-JNJ5].

23. *Alexa Tells 10-Year-Old Girl to Touch Live Plug with Penny*, BBC (Dec. 28, 2021), https://www.bbc.com/news/technology-59810383 [https://perma.cc/ME2W-ZGUE].

24. *See, e.g.*, Peter Henderson, Tatsunori Hashimoto & Mark Lemley, *Where's the Liability in Harmful AI Speech?*, 3 J. FREE SPEECH L. 589 (2023); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315, 1319 (2020); Daniel J. Solove & Hideyuki Matsumi, *AI, Algorithms, and Awful Humans*, 92 FORDHAM L. REV. 1923 (2024); Bobby Chesney & Danielle Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1772–73 (2019); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851, 1861 (2019); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017) (book review); Sylvia Lu, Note, *Data Privacy, Human Rights, and Algorithmic Opacity*, 110 CALIF. L. REV. 2087, 2105 (2022); Sarah M.L. Bender, Note, *Algorithmic Elections*, 121 MICH. L. REV. 489, 491 (2022).

25. *See, e.g.*, Ifeoma Ajunwa, *Automated Video Interviewing as the New Phrenology*, 36 BERKELEY TECH. L.J. 1173, 1190–94 (2021); Sonja B. Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043, 1104 (2019); Shima Baradaran, *Race, Prediction, and Discretion*, 81 GEO. WASH. L. REV. 157 (2013); Andrew Guthrie Ferguson, *Illuminating Black Data Policing*, 15 OHIO ST. J. CRIM. L. 503, 516 (2018); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016); Sandra G. Mayson, *Bias in, Bias out*, 128 YALE L.J. 2218, 2251 (2019); Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007 (2022).

26. Eugene Volokh, *Large Libel Models? Liability for AI Output*, 3 J. FREE SPEECH L. 489, 493 (2023).

27. Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1339 (2019); M. Ryan Calo, *Robots and Privacy*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 187 (Patrick Lin, Keith Abney & George A Bekey eds., 2012).

28. Chesney & Citron, *supra* note 24, at 1774.

from negligence to reckless enablement of a crime.[29] Avoiding a myopic focus on a plaintiff's case, some academics have explored the possible defenses AI companies could use to quash lawsuits against them.[30] Beyond litigation, scholars have proposed policy reforms to address AI governance issues, such as issuing revocable licenses for AI products or regulating AI's ability to give professional advice.[31]

One scholar, Bryan Choi, has raised the possibility of treating software engineers as professionals for the purposes of malpractice liability, but he stops short of advocating for proactive professionalization as a solution.[32] This approach captures some of the value of true professionalization, though it leaves much more on the table. Without the top-down establishment of self-governing institutions and the bottom-up effects of culture change, shifting the standard of care on its own is unlikely to reform AI development practices. Further, Choi is actually skeptical of the appropriateness of a professional standard of care for AI engineers, as opposed to traditional software developers, pointing to the field's nascency and lack of consensus on standards.[33] However, those are the very reasons for professionalization: harnessing latent expertise and good intentions to force the discipline to mature.

---

29. *See, e.g.*, Charlotte A. Tschider, *Humans Outside the Loop*, 26 YALE J.L. & TECH. 324 (2024); Selbst, *supra* note 24; Marguerite E. Gerstner, Comment, *Liability Issues with Artificial Intelligence Software*, 33 SANTA CLARA L. REV. 239, 246–54 (1993). *See generally* Michael L. Rustad & Thomas H. Koenig, *The Tort of Negligent Enablement of Cybercrime*, 20 BERKLEY TECH. L.J. 1553 (2005); Mark A. Geistfeld, *A Roadmap for Autonomous Vehicles: State Tort Liability, Automobile Insurance, and Federal Safety Regulation*, 105 CALIF. L. REV. 1611, 1612 (2017); Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Iria Giuffrida, *Liability for AI Decision-Making: Some Legal and Ethical Considerations*, 88 FORDHAM L. REV. 439 (2019).

30. *See, e.g.*, Amy L. Stein, *Assuming the Risks of Artificial Intelligence*, 102 B.U. L. REV. 979, 986–96 (2022); Amy B. Cyphert & Jena T. Martin, *"A Change is Gonna Come:" Developing a Liability Framework for Social Media Algorithmic Amplification*, 13 U.C. IRVINE L. REV. 155, 186 (2022).

31. Gianclaudio Malgieri & Frank Pasquale, *From Transparency to Justification: Toward* Ex Ante *Accountability for AI* 2 (Brooklyn L. Sch., Working Paper No. 712; Brussels Priv. Hub Working Paper, Paper No. 33, 2022), https://ssrn.com/abstract=4099657 [https://perma.cc/9YY7-MWN4]; Margot E. Kaminski, *The Developing Law of AI: A Turn to Risk Regulation*, DIGIT. SOC. CONT.: A LAWFARE PAPER SERIES, Apr. 2023; Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119 COLUM. L. REV. 369, 406 (2019); Alicia Solow-Niederman, *Administering Artificial Intelligence*, 93 S. CAL. L. REV. 633 (2020); Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429 (2023); Frank Pasquale, *Data-Informed Duties in AI Development*, 119 COLUM. L. REV. 1917 (2019); Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018); Claudia E. Haupt, *Artificial Professional Advice*, 21 YALE J.L. & TECH. 55, 72–76 (2019); Andrew D. Selbst & Solon Barocas, *Unfair Artificial Intelligence: How FTC Intervention Can Overcome the Limitations of Discrimination Law*, 171 U. PA. L. REV. 1023 (2023) [hereinafter Selbst & Barocas, *Unfair Artificial Intelligence*].

32. Bryan H. Choi, *Software as a Profession*, 33 HARV. J.L. & TECH. 557, 603–09 (2020).

33. *See generally* Bryan H. Choi, *AI Malpractice*, 73 DEPAUL L. REV. 301 (2024).

Overall, the solutions proposed in existing scholarship fall into two categories of interventions: substantive regulation and legal liability. In the absence of ideal market behavior, some call on government to take on the Sisyphean-seeming task of regulating AI. While these proposals are laudable, the devil is in the details of AI engineering,[34] and governments lack the expertise and nimbleness to meaningfully regulate AI with the level of particularity necessary to influence its design—worse still, they are prone to tech favoritism, leaving them vulnerable to regulatory capture.[35] Others look to litigation, but it is also limited in its ability to hold tech accountable given the way AI complicates questions of harm, standing, and Section 230.[36] However vital to a holistic solution, these types of intervention are insufficient on their own to tackle the intractable AI problem.

Professionalization is up to the task. It has distinct advantages over substantive regulation and legal action, and it is familiar from the fields of medicine, accounting, law, and engineering. Professionalization's formal institutions can minimize the risk of technical errors, while its power to transform an individual engineer's desire to do good into a culture of social responsibility can minimize the risk of ethical errors—orienting a discipline toward serving the public interest over customer priorities. It capitalizes on the expertise and agility of professionals, rather than bureaucrats, to set and update legally enforceable technical and ethical standards. It also creates a new vehicle for accountability, which bypasses the barriers holding tort law hostage: professional tribunals investigating and disciplining AI engineers accused of malpractice by their peers or the public—enlisting the industry to police itself.

Through information sharing and mandatory, enforceable practice guidelines, professionalization guards against the ignorant or willful disregard for technical standards. In this way, it raises the field's ability to deliver quality service, earning the public's trust and regard. Through its ability to organize a disparate group of engineers as a formal discipline, with the imprimatur of legitimacy, professionalization can galvanize the altruistic undercurrent motivating many AI engineers into a culture of social responsibility.

Professionalization can also hold its own against market pressures, emboldening AI engineers to pump the brakes on the AI arms race, refusing company directives to build more, faster, if doing so risks malpractice. It can also break down the walled gardens companies erect to withhold information about their AI systems. This would allow AI engineers to share

34. *See infra* Section I.A.
35. *See infra* Section I.C.
36. *See infra* Section I.C.1.

information with professionals outside of their companies to identify risks and design solutions *for the common good*.

Professionalization forces AI engineers to consider the public's interest. Other scholars have called for similar duties on the technology industry, such as information fiduciary duties,[37] especially a duty of loyalty.[38] Some have urged judges to hold software engineers to a professional standard of care in lawsuits.[39] This Article's proposal is consistent with that body of work, but it goes further. Professionalization creates more than an abstract duty: it produces a sophisticated system of institutions and policies that define, execute, and enforce that duty.

Although this Article provides concrete policy recommendations, its major contribution is to inject a new perspective on AI governance into the discourse. Some proposals focus on regulating the behavior of AI companies, ignoring the body of scholarship warning that imposing high compliance costs can entrench already dominant players.[40] Others focus on regulating the outputs of AI or the final product, which comes too little too late, appraising a technology that defies inspection for risks that, if found, are nearly impossible to undo.[41] The law has an important role to play from the very beginning of the AI engineering process, where it has the best opportunity to build in safety, security, and trustworthiness by design. This Article argues that the most effective pressure point to influence the AI engineering process from its inception is the individual AI engineer.

Further, this Article interacts with literature beyond AI. The *perils* of professionalization are well documented in legal scholarship.[42] Yet, even professionalization's harshest critics concede it is necessary for services that require an immense amount of expertise, the quality of which is difficult for customers to evaluate, in contexts that have broader implications for society.[43] AI more than fits the bill.

This Article makes the novel proposition that AI engineers should professionalize. In doing so, it makes three distinct contributions. In Part I, the Article presents the multifaceted AI problem—a Gordian knot of high stakes uses, inscrutable technology, irresponsible engineering practices, misaligned market incentives, and corporate secrecy or subterfuge. It also

37. Jack M. Balkin, *The Fiduciary Model of Privacy*, 134 HARV. L. REV. F. 11, 14 (2020).

38. Woodrow Hartzog & Neil Richards, *The Surprising Virtues of Data Loyalty*, 71 EMORY L.J. 985, 989 (2022).

39. *See* Choi, *supra* note 32, at 563 n.25.

40. *See infra* Section I.C.2.

41. *See infra* Section I.C.1.

42. *See, e.g.*, Rebecca Haw Allensworth, *Foxes at the Henhouse: Occupational Licensing Boards Up Close*, 105 CALIF. L. REV. 1567, 1570 (2017); Aaron Edlin & Rebecca Haw, *Cartels by Another Name: Should Licensed Occupations Face Antitrust Scrutiny?*, 162 U. PA. L. REV. 1093, 1095–96, 1108 (2014).

43. *See infra* Part III.

explains why substantive regulation and traditional forms of legal action are insufficient, on their own, to resolve the AI problem. Part II introduces the novel solution of professionalizing AI engineers. It identifies three distinct advantages of professionalization: (1) it forces the AI industry to prioritize the public interest in building these powerful technologies; (2) it empowers individual AI engineers, thereby breaking Big Tech's stranglehold on the AI ecosystem; and (3) it overcomes many of the roadblocks limiting the effectiveness of substantive regulation and traditional litigation. Part III explores what this solution looks like on the ground by first enumerating the various steps required to professionalize the field of AI engineering and then providing recommendations to circumvent the risks professionalization may present, such as anticompetitive practices.

I. THE INTRACTABLE AI PROBLEM

Alongside its benefits to society, AI comes with a generous helping of harms. This Part introduces the "AI problem," or the notion that AI, however promising, presents new and alarming risks that neither the market, nor government, nor courts are well-suited to address on their own. First, it illustrates the many ways AI threatens society, from individual well-being to public institutions, and explains how these harms are caused by decisions made in the AI engineering process. Then, it identifies the characteristics about the AI marketplace that exacerbate these harms, ensuring their continuation without intervention. Finally, it explains why our traditional methods of minimizing harms through regulation and litigation are inadequate to curb AI harms.

A. The Source of Harmful AI

Technology is never perfectly safe, but AI's unique characteristics present unique risks. This section explains how AI's raw power and ubiquity raises the stakes of the AI problem. It then describes how AI harms manifest and traces their source back to the AI engineering process. Ultimately, it demonstrates that harmful AI comes from irresponsible engineering.

1. A Pressing Problem

Although AI is hardly novel,[44] it jumped to the forefront of the public's imagination with the explosion of ChatGPT onto the scene in the fall of

44. *See* Rockwell Anyoha, *The History of Artificial Intelligence*, SCI. NEWS (Aug. 28, 2017), https://sitn.hms.harvard.edu/flash/2017/history-artificial-intelligence/ [https://perma.cc/LC3N

2022.[45] One day, we were all laughing at Siri's awkward and distinctly inhuman responses to prompts[46] and AI was largely[47] operating in the shadows.[48] The next, we were embracing a world order in which judges use AI to write opinions[49] and mayors distribute deep fakes of themselves speaking different languages to their constituents.[50]

However, just as the sharpest knife is both the most effective and the most hazardous, so too is AI's immense power both a boon and a threat. Now that AI is everywhere, its potential risks are now woven into the fabric of our day-to-day lives, creating a minefield littered with opportunities for AI to cause harm. Preserving AI's potential to help humanity while mitigating its ability to cause harm demands an understanding of when AI helps and when it hurts.

AI's value lies in its ability to both help humans do things better and do things humans cannot do at all. Although it is difficult to draw the line between traditional software and artificial intelligence, domestic and international regulation have already taken a stab at the problem.[51] At a

-ELQD] (explaining that AI has been around for decades); A.M. Turing, *Computing Machinery and Intelligence*, 59 MIND 433 (1950) (foreshadowing the use of anthropomorphized AI in human-decision making in 1950).

45. Bernard Marr, *A Short History of ChatGPT: How We Got to Where We Are Today*, FORBES (May 19, 2023, 1:14 AM), https://www.forbes.com/sites/bernardmarr/2023/05/19/a-short-history-of-chatgpt-how-we-got-to-where-we-are-today/ [https://perma.cc/4P49-AZW5].

46. Page Laubheimer & Raluca Budiu, *Intelligent Assistants: Creepy, Childish, or a Tool? Users' Attitudes Toward Alexa, Google Assistant, and Siri*, NN/G (Aug. 5, 2018), https://www.nngroup.com/articles/voice-assistant-attitudes/ [https://perma.cc/H453-3Y6F].

47. *See Deep Blue*, IBM, https://www.ibm.com/history/deep-blue [https://perma.cc/8WFK-H3H5] (chronicling the history of IBM's AI, Deep Blue, which beat a chess grandmaster); *Google AI Defeats Human Go Champion*, BBC (May 25, 2017), https://www.bbc.com/news/technology-40042581 [https://perma.cc/G7CL-EPLS] (covering the story of Google's AlphaGo AI system beating the world champion in the game of Go).

48. *See, e.g.*, Kerry Breen, Brook Silva-Braga & Greg Mirman, *Some Experts Push for Transparency, Open Sourcing in AI Development*, CBS NEWS (Dec. 16, 2023, 9:57 AM), https://www.cbsnews.com/news/artificial-intelligence-yan-lecun-meta-transparency-open-sourcing/ [https://perma.cc/F77K-N4JA] (explaining that since Facebook hired Yann LeCun in 2013, long before ChatGPT, the AI engineer has been embedding AI in systems that recommend friends, optimize ads, and automatically censor posts that violate the platform's roles).

49. *See* Melley, *supra* note 18.

50. Anthony Izaguirre, *Eric Adams' Revelation that He Uses AI to Speak in Mandarin Stirs Outcry: 'The Mayor is Making Deep Fakes of Himself'*, FORTUNE (Oct. 17, 2023, 2:37 PM), https://fortune.com/2023/10/17/new-york-city-mnayor-eric-adams-uses-ai-to-speak-mandarin/ [https://perma.cc/Z5MN-6J3K].

51. Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023) (defining artificial intelligence as "a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments"); Regulation 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonized Rules on Artificial Intelligence, art. 3, 2024 O.J. (L 12.7) (EU) [hereinafter *AI Act*] (defining an AI system as "a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments").

minimum, AI is not a single product or even a category of recognizable products; it is a *scientific field* developing computer systems that make intelligent decisions.[52] In other words, AI translates complex mathematical formulas into code and harnesses the processing power of computers to actually *do* the math. Because a computer's processing power far outstrips the human brain, AI systems are capable of accomplishing much more than even the smartest among us.[53] As computers,[54] math,[55] and the availability of data[56] improved over the years, AI's power grew.

Today, AI decisions can appear to us as an answer to a search query,[57] a credit score,[58] a conversation reply,[59] a cartoon generated image,[60] a fake voice memo,[61] or a robot dog rolling over.[62] These decisions are really just predictions of the right answer based on applying statistical models to

52. CHRISTOPHER MANNING, STAN. U. HUMAN-CENTERED A.I., ARTIFICIAL INTELLIGENCE DEFINITIONS 1 (2020), https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf [https://perma.cc/PE3X-Q5U2].

53. Sara Reardon, *Artificial Neurons Compute Faster than the Human Brain*, NATURE (Jan. 26, 2018), https://www.nature.com/articles/d41586-018-01290-0 [https://perma.cc/4US8-HWUW]; Selbst, *supra* note 24, at 1319 ("Instead of seeking to replicate human capabilities such as driving, they often seek to go beyond human capabilities, recognizing and modeling patterns too complex for humans to process and making decisions in ways humans would not recognize.").

54. Intel AI, *The Rise in Computing Power: Why Ubiquitous Artificial Intelligence Is Now a Reality*, FORBES (May 28, 2019, 11:11 AM), https://www.forbes.com/sites/intelai/2018/07/17/the-rise-in-computing-power-why-ubiquitous-artificial-intelligence-is-now-a-reality/ [https://perma.cc/9GJA-CQDF].

55. Emerging Technology from the arXiv, *The Extraordinary Link Between Deep Neural Networks and the Nature of the Universe*, MIT TECH. REV. (Sept. 9, 2016), https://www.technologyreview.com/2016/09/09/157625/the-extraordinary-link-between-deep-neural-networks-and-the-nature-of-the-universe/ [https://perma.cc/7A3Y-YQSS]; Matthew Hutson, *DeepMind AI Invents Faster Algorithms to Solve Tough Maths Puzzles*, NATURE (Oct. 5, 2022), https://www.nature.com/articles/d41586-022-03166-w [https://perma.cc/W57L-5AP4].

56. Yoshua Bengio, *Springtime for AI: The Rise of Deep Learning*, SCI. AM. (June 1, 2016), https://www.scientificamerican.com/article/springtime-for-ai-the-rise-of-deep-learning/ [https://perma.cc/PZC3-7RWQ].

57. Yusuf Mehdi, *Reinventing Search with a New AI-Powered Microsoft Bing and Edge, Your Copilot for the Web*, MICROSOFT (Feb. 7, 2023), https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/ [https://perma.cc/8KQR-Q5FD].

58. Sean Michael Kerner, *How AI 'Data Drift' May Have Caused the Equifax Credit Score Glitch*, VENTUREBEAT (Aug. 16, 2022, 8:59 AM), https://venturebeat.com/data-infrastructure/did-data-drift-in-ai-models-cause-the-equifax-credit-score-glitch/ [https://perma.cc/64XZ-R6ZD].

59. *ChatGPT Can Now See, Hear, and Speak*, OPENAI (Sept. 25, 2023), https://openai.com/blog/chatgpt-can-now-see-hear-and-speak [https://perma.cc/9UNY-9YXE].

60. *See generally* Yang Chen, Yu-Kun Lai & Yong-Jin Liu, *CartoonGAN: Generative Adversarial Networks for Photo Cartoonization*, *in* 2018 IEEE/CVF CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION 9465 (2018), https://openaccess.thecvf.com/content_cvpr_2018/papers/Chen_CartoonGAN_Generative_Adversarial_CVPR_2018_paper.pdf [https://perma.cc/M6WJ-RWWX].

61. *The Best Fake Voice Generator*, SPEECHIFY, https://speechify.com/fake-voice-generator/ [https://perma.cc/3KRG-A8Y7].

62. Melissa Heikkilä, *This Robot Dog Just Taught Itself to Walk*, MIT TECH. REV. (July 18, 2022), https://www.technologyreview.com/2022/07/18/1056059/robot-dog-ai-reinforcement/ [https://perma.cc/8RBP-XMEC].

inputted information—after having read the full works of Shakespeare and Taylor Swift's discography, AI can predict what a Taylor Swift song written in iambic pentameter would look like.

Given the promise of so much power, AI has seeped into every corner of our lives. AI can offer life advice[63] and serve as a virtual wingman.[64] It helps us work faster and produce better work;[65] not insignificant in a world in which people are working longer hours[66] and getting less quality sleep.[67] Beyond raw productivity, AI is also bridging gender, class and ethnic divides by assisting in salary negotiations[68] and seamlessly translating languages, including some that do not have widely used writing systems.[69] Companies are turning to AI for content moderation[70] because it may be comparable to and substantially cheaper than human moderators,[71] which could spare countless people the trauma of filtering the internet's vilest content.[72]

But AI can be flaky. When it is not helpful, it is powerfully harmful, and it is not always easy to predict which it will be for any given task. For example, the field of medicine has long used AI for diagnoses, genetic indexes, and personalized medicine.[73] But, as is true for AI across the board, it is better at some tasks than others. While AI is excellent at diagnosing

63. Nico Grant, *Google Tests an A.I. Assistant That Offers Life Advice*, N.Y. TIMES (Aug. 16, 2023), https://www.nytimes.com/2023/08/16/technology/google-ai-life-advice.html [https://perma.cc/YH4W-YUYU].

64. *See* Herrman, *supra* note 13.

65. Matt Albasi, *New Report: The State of AI in PR 2024*, MUCK RACK (Jan. 4, 2024), https://muckrack.com/blog/2024/01/04/state-of-ai-in-pr-2024 [https://perma.cc/VTT9-JMEF] (explaining that 74% of public relations professionals said it improved the quality of their work and 89% said it helped them complete their projects faster).

66. Kapo Wong, Alan H.S. Chan & S.C. Ngan, *The Effect of Long Working Hours and Overtime on Occupational Health: A Meta-Analysis of Evidence from 1998 to 2018*, INT'L J. ENV'T RSCH. & PUB. HEALTH, June 13, 2019, at 1 (explaining the adverse health effects of working long hours).

67. Jagdish Khubchandani & James H. Price, *Short Sleep Duration in Working American Adults, 2010–2018*, 45 J. CMTY. HEALTH 219, 222 (2020).

68. Ina Fried & Ryan Heath, *1 Big Thing: AI's Next Target is Salary Negotiations*, AXIOS (Aug. 25, 2023), https://www.axios.com/newsletters/axios-ai-plus-7d05746c-4004-4de1-9dea-c9604598d7a6.html [https://perma.cc/LAS7-M7FD].

69. Alison Snyder, *AI's Language Gap*, AXIOS (Sept. 8, 2023), https://www.axios.com/2023/09/08/ai-language-gap-chatgpt [https://perma.cc/UC5A-5Q6N].

70. Casey Newton, *OpenAI Wants to Moderate Your Content*, PLATFORMER (Aug. 15, 2023), https://www.platformer.news/openai-wants-to-moderate-your-content [https://perma.cc/Y6B4-ZX76].

71. Fabrizio Gilardi, Meysam Alizadeh & Maël Kubli, *ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks*, PROC. NAT'L. ACAD. SCIENCES, July 18, 2023, at e2305016120.

72. Casey Newton, *Facebook Will Pay $52 Million in Settlement with Moderators Who Developed PTSD on the Job*, THE VERGE (May 12, 2020, 3:39 PM), https://www.theverge.com/2020/5/12/21255870/facebook-content-moderator-settlement-scola-ptsd-mental-health [https://perma.cc/9XWJ-AFWJ].

73. *See, e.g.*, W. Nicholson Price II, *Regulating Black-Box Medicine*, 116 MICH. L. REV. 421, 425–26 (2017).

some cancers based on radiology images,[74] it struggles to make other kinds of initial diagnoses.[75] It excels at standardized testing for medical students,[76] including on open-ended clinical reasoning questions,[77] but fails to produce medically accurate and useful notes based on electronic health records.[78] Success at one task is not a great indicator of success at another, even in the same industry, such as medicine. General purpose systems like ChatGPT exacerbate this problem.[79] Curious users are bound to test the seemingly endless limits of the technology, inevitably venturing into use cases for which AI is not a good fit. Worse still, users are bad at detecting harmful AI and so cannot avoid the mines in the minefield.[80]

On one hand, failing to use AI where it can be helpful risks missing an opportunity to serve the public. For example, the United Kingdom National Health Service approved an AI system capable of diagnosing patients with ninety-three percent accuracy across eight of the most common mental disorders as the equivalent of a Class II medical device[81]—which the Food

74. Hyo-Eun Kim et al., *Changes in Cancer Detection and False-Positive Recall in Mammography Using Artificial Intelligence: A Retrospective, Multireader Study*, 2 LANCET DIGIT. HEALTH e138 (2020) (summarizing a large retrospective study in which a deep learning algorithm was developed and validated with 170,230 mammography examinations collected from five institutions in South Korea, the United States, and the UK, showing a "better diagnostic performance in breast cancer detection compared with radiologists").

75. Arya Rao et al., *Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study*, J. MED. INTERNET RSCH., 2023, at e48659 (finding that ChatGPT achieved 71.7% accuracy across all thirty-six clinical vignettes, with the highest performance at making a final diagnosis and the lowest performance at making an initial differential diagnosis).

76. HARSHA NORI, NICHOLAS KING, SCOTT MAYER MCKINNEY, DEAN CARIGNAN & ERIC HORVITZ, CAPABILITIES OF GPT-4 ON MEDICAL CHALLENGE PROBLEMS 1 (2023), https://arxiv.org/pdf/2303.13375 [https://perma.cc/9BWM-DYLW] (reporting that ChatGPT can successfully handle multiple-choice questions on the United States Medical License Examination, which doctors must pass in order to practice medicine).

77. Eric Strong et al., *Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations*, 183 JAMA INTERNAL MED. 1028, 1028 (2023) (finding that the AI model, "on clinical reasoning skills[,] . . . outperformed students on creating a problem list by 16 points").

78. SCOTT L. FLEMING ET AL., MEDALIGN: A CLINICIAN-GENERATED DATASET FOR INSTRUCTION FOLLOWING WITH ELECTRONIC MEDICAL RECORDS 1 (2023), https://arxiv.org/pdf/2308.14089 [https://perma.cc/T6FB-7X8V].

79. Where once, AI systems were "specialized tools" designed to serve specific functions, such as diagnosing breast cancer or playing chess, today the field of AI is shifting to produce more general-purpose AI systems. Haomiao Huang, *How ChatGPT Turned Generative AI into an "Anything Tool,"* ARS TECHNICA (Aug. 23, 2023 7:30 AM), https://arstechnica.com/ai/2023/08/how-chatgpt-turned-generative-ai-into-an-anything-tool [https://perma.cc/MV82-8ZMY] ("Until recently, AI models were specialized tools.").

80. *See* SAMIR PASSI & MIHAELA VORVOREANU, AETHER, OVERRELIANCE ON AI: LITERATURE REVIEW 10 (2022), https://www.microsoft.com/en-us/research/uploads/prod/2022/06/Aether-Overreliance-on-AI-Review-Final-6.21.22.pdf [https://perma.cc/TD2U-C746] ("Users find it difficult to evaluate AI's performance and to understand how AI impacts their decisions. For instance, users often overestimate system accuracy and do not realize when they have ceded control to the AI." (citations omitted)).

81. Ben Carrington, *AI Mental Health Chatbot That Predicts Disorders Becomes First in World to Gain Class IIa UKCA Medical Device Status*, LIMBIC (Jan. 17, 2023), https://limbic.ai/blog/class-ii-a [https://perma.cc/Y98P-FG4S].

& Drug Administration ("FDA") classifies as medium-risk.[82] Amidst a global mental health crisis coming out of the pandemic,[83] this kind of tool can get patients the healthcare they need faster.[84] On the other hand, using AI in the wrong way risks undermining patient care and health outcomes.

2. A Range of Harms

Today, harmful AI systems flourish, capitalizing on society's enchantment with the technology and its inability to discern helpful AI from harmful AI. Although measuring accurate incident rates for diffuse AI harms is notoriously hard, studies have estimated that the rate of harmful AI increased by a factor of twenty-six since 2012,[85] ranging from victimizing minors and causing hospitalizations[86] to running smear campaigns[87] and perpetuating structural oppression. And this is just the beginning.

Scams are timeless. But with AI, scams are being perpetrated with an unprecedented degree of sophistication and magnitude. AI powered voice impersonation technology has been used to scam a company out of thirty-five million[88] and demand ransom from a mother in her daughter's voice.[89] Writing convincing phishing emails can take a team of security professionals sixteen hours—ChatGPT can do it in five minutes.[90] AI has

82. *Classify Your Medical Device*, FDA (Feb. 7, 2020), https://www.fda.gov/medical-devices/overview-device-regulation/classify-your-medical-device [https://perma.cc/Q7EL-SA4Z] (classifying Class II medical devices as medium risk).

83. *The Impact of COVID-19 on Mental Health Cannot be Made Light Of*, WORLD HEALTH ORG. (June 16, 2022), https://www.who.int/news-room/feature-stories/detail/the-impact-of-covid-19-on-mental-health-cannot-be-made-light-of [https://perma.cc/BM5Q-NRX6].

84. *See, e.g.*, Carrington, *supra* note 81.

85. STAN. UNIV. HUMAN-CENTERED A.I., ARTIFICIAL INTELLIGENCE INDEX REPORT 2023, at 133 (2023), https://aiindex.stanford.edu/wp-content/uploads/2023/04/HAI_AI-Index-Report_2023.pdf [https://perma.cc/E4HU-UUMH].

86. *See, e.g.*, Sarah Perez, *Several Popular AI Products Flagged as Unsafe for Kids by Common Sense Media*, TECHCRUNCH (Nov. 16, 2023, 8:06 AM), https://techcrunch.com/2023/11/16/several-popular-ai-products-flagged-as-unsafe-for-kids-by-common-sense-media/ [https://perma.cc/6SNQ-W7L3]; Jasper Jolly, *Amazon Robot Sets Off Bear Repellant, Putting 24 Workers in Hospital*, THE GUARDIAN (Dec. 6, 2018, 8:00 AM), https://www.theguardian.com/technology/2018/dec/06/24-us-amazon-workers-hospitalised-after-robot-sets-off-bear-repellent [https://perma.cc/LS85-3GC9].

87. Byron Kaye, *Australian Mayor Readies World's First Defamation Lawsuit over ChatGPT Content*, REUTERS (Apr. 5, 2023, 2:52 PM), https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/ [https://perma.cc/9S8N-RG57]; *see* Volokh, *supra* note 26, at 493.

88. Tim Wu, *In Regulating A.I., We May Be Doing Too Much. And Too Little.*, N.Y. TIMES (Nov. 7, 2023), https://www.nytimes.com/2023/11/07/opinion/biden-ai-regulation.html [https://perma.cc/QMK3-7Z3X].

89. Ricky Nave, *6 Cases of AI Crime: Kidnapping Scams, Suicide & Class Action Lawsuits*, AXIOM ALPHA, https://axiomalpha.com/6-cases-of-ai-crime-kidnapping-scams-suicide-class-action-lawsuits/ [https://perma.cc/N322-L7ZC].

90. Stephanie Carruthers, *AI vs. Human Deceit: Unravelling the New Age of Phishing Tactics*, SEC. INTEL. (Oct. 24, 2023), https://securityintelligence.com/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics/ [https://perma.cc/9EK9-U4R2].

also supercharged the thriving market for fake reviews, which today comprise thirty to forty percent of total online reviews, used to mislead customers and sink competitors.[91] Additionally, AI is itself a honeypot for hackers.[92] It can be "hypnotized" into regurgitating personal user information or confidential business information it has collected.[93] There is even a derivative market for exploits to jailbreak AI so anyone can join the fun.[94]

Anthropomorphized AI can be a wolf in sheep's clothing. Humans trust AI more when it is personified,[95] when they are primed to believe it is caring,[96] and when it offers detailed but spurious explanations for its recommendations.[97] If charismatic AI told us to jump off a bridge, we might. Alexa tried to persuade a ten year old to play the "penny game," which entailed touching a partially inserted live plug with a coin, but the child wisely refused, alerting her mother to the exchange.[98] An environmentalist killed himself, convinced by AI that doing so would be the best way to help his cause.[99] It can subversively change minds on important issues such as gun control.[100]

---

91. *See* Geoffrey A. Fowler, *Those 10,000 5-Star Reviews Are Fake. Now They'll Also Be Illegal.*, WASH. POST (June 30, 2023, 5:59 PM), https://www.washingtonpost.com/technology/2023/06/30/fake-reviews-online-ftc/ [https://perma.cc/GX83-GVKG].

92. *See* Joseph Clark, *AI Security Center to Open at National Security Agency*, U.S. DEP'T OF DEF. (Sept. 28, 2023), https://www.defense.gov/News/News-Stories/Article/Article/3541838/ai-security-center-to-open-at-national-security-agency/ [https://perma.cc/A26T-XVPW].

93. *See* Chenta Lee, *Unmasking Hypnotized AI: The Hidden Risks of Large Language Models*, SEC. INTEL. (Aug. 8, 2023), https://securityintelligence.com/posts/unmasking-hypnotized-ai-hidden-risks-large-language-model [https://perma.cc/56NF-99BE]; *see also* Michelle Drolet, *10 Ways Cybercriminals Can Abuse Large Language Models*, FORBES (June 30, 2023, 8:45 AM), https://www.forbes.com/sites/forbestechcouncil/2023/06/30/10-ways-cybercriminals-can-abuse-large-language-models/ [https://perma.cc/AN22-PTW2]; Zoë Schiffer & Casey Newton, *Amazon's Q Has 'Severe Hallucinations' and Leaks Confidential Data in Public Preview, Employees Warn*, PLATFORMER (Dec. 1, 2023), https://www.platformer.news/p/amazons-q-has-severe-hallucinations [https://perma.cc/AT2Y-8F2W].

94. *See* Rachel Metz, *Jailbreaking AI Chatbots Is the Tech Industry's New Pastime*, BLOOMBERG L. (Apr. 8, 2023, 9:00 AM), https://news.bloomberglaw.com/private-equity/jailbreaking-ai-chatbots-is-the-tech-industrys-new-pastime [https://perma.cc/5F9H-B9NQ].

95. Lorenzo Cominelli et al., *Promises and Trust in Human-Robot Interaction*, 11 SCI. REPS. 9687 (2021); Ella Glikson & Anita Williams Woolley, *Human Trust in Artificial Intelligence: Review of Empirical Research*, 14 ACAD. MGMT. ANNALS 627, 627 (2020); Woodrow Hartzog, *Unfair and Deceptive Robots*, 74 MD. L. REV. 785, 787 (2015).

96. Pat Pataranutaporn, Ruby Liu, Ed Finn & Pattie Maes, *Influencing Human–AI Interaction by Priming Beliefs About AI Can Increase Perceived Trustworthiness, Empathy and Effectiveness*, 5 NATURE MACH. INTEL. 1076, 1082 (2023).

97. *See* PASSI & VORVOREANU, *supra* note 80, at 9.

98. BBC, *supra* note 23.

99. Xiang, *supra* note 19.

100. For instance, authors of one study report that AI-generated messages were at least as persuasive as human-generated messages across all topics, and on a smoking ban, gun control, carbon tax, an increased child tax credit, and a parental leave program, participants became "significantly more supportive" of the policies when reading AI-produced texts. HUI BAI, JAN G. VOELKEL, JOHANNES C.

The risks are more pronounced when vulnerable individuals seek help from AI. Google's AI safety experts have said that users could experience "diminished health and well-being" if they took life advice from AI.[101] When patients sought medical support from the National Eating Disorder Association's AI bot, it began encouraging them to starve themselves.[102] AI has even gone so far as to teach people to cut themselves,[103] provide recipes to cook poisonous mushrooms,[104] and allow the purchase, by children, of chemicals known for their use in suicides.[105] And AI's manipulative abilities have lasting power; in one study, physicians continued to follow AI's advice *even after being told the AI was wrong*.[106]

Societal harms are felt most often and severely by its most marginalized groups—AI is no different.[107] "Deepfake" intimate images have always targeted women far more than other populations.[108] Where once, this form of hateful misogyny required some degree of skill, AI obliterated the barriers to entry, allowing for both the creation of increasingly convincing fake images and their widespread distribution with little to no effort.[109] Translation models today are trained on and therefore useful for English and Chinese data, further marginalizing the "6 billion native speakers of the world's more than 7,000 other languages."[110]

EICHSTAEDT & ROBB WILLER, ARTIFICIAL INTELLIGENCE CAN PERSUADE HUMANS ON POLITICAL ISSUES 2 (2023); Shangbin Feng, Chan Young Park, Yuhan Liu & Yulia Tsvetkov, *From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models*, *in* 1 PROCEEDINGS OF THE 61ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 11737 (2023) (arguing that the major LLMs available today exhibit political biases ranging across the spectrum of beliefs, which could impact the public by, for example, denying a patient seeking medical advice about a pregnancy information about abortions).

101. Grant, *supra* note 63.

102. PSYCHIATRIST.COM, *supra* note 20.

103. OPENAI, GPT-4 SYSTEM CARD 47–50 (2023), https://cdn.openai.com/papers/gpt-4-system-card.pdf [https://perma.cc/58P2-V5XT]; *see also* Henderson et al., *supra* note 24, at 605.

104. Thomas Germain, *'Benefits of Slavery:' Google's AI Search Gives Ridiculous and Wrong Answers*, GIZMODO (Aug. 22, 2023, 10:16 PM), https://gizmodo.com/google-search-ai-answers-slavery-benefits-1850758631?rev=1692644130500 [https://perma.cc/ZXA3-KW27].

105. Jonathan Stempel, *Judge Dismisses Lawsuit Claiming Amazon Sold 'Suicide Kits' to Teenagers*, REUTERS (June 28, 2023, 12:13 PM), https://www.reuters.com/legal/judge-dismisses-lawsuit-claiming-amazon-sold-suicide-kits-teenagers-2023-06-28/ [https://perma.cc/GD8L-PGNP].

106. Lucía Vicente & Helena Matute, *Humans Inherit Artificial Intelligence Biases*, 13 SCI. REPS. 15737 (2023).

107. Ethan Zuckerman, *Want to Stop Harmful Tech? Just Say No*, PROSPECT (July 19, 2023), https://www.prospectmagazine.co.uk/ideas/technology/62163/want-to-stop-harmful-tech-just-say-no [https://perma.cc/YV4R-ED3D] ("But AI is already harming humans. And it's disproportionately harming people of colour, welfare recipients and the unhoused. To imagine AI inflicting harm on the privileged requires a leap of faith in the power of technology. Understanding existing harm to the vulnerable does not . . . .").

108. Anne Pechenik Gieseke, Note, *"The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography*, 73 VAND. L. REV. 1479, 1482 (2020).

109. *See* Chesney & Citron, *supra* note 24, at 1774.

110. Snyder, *supra* note 69.

Models also further structural harm when AI engineers fail to account for underlying biases in training data.[111] Google's AI generated search results have defended slavery and genocide, spewing language used historically to justify these atrocities.[112] Facial recognition systems are trained on data overrepresenting Caucasian[113]—often male—faces, making them potentially ten to one hundred times worse at identifying people of color,[114] which has already led to several instances of people of color being misapprehended for crimes they did not commit.[115] With the current shift toward general purpose AI systems, these harms are likely to trickle downstream, corrupting every technology, institution, or decision they touch with problematic decision-making.[116]

In the future, without intervention, these harms will be joined by harms of a different kind: eroding trust in institutions.[117] Deepfakes threaten democracy.[118] Already, the Republican National Committee responded to President Biden's re-election announcement with an AI-generated video[119]—more subversive uses of AI in political campaigns are just around the corner.[120] Specifically, experts fear disinformation campaigns used to

111. *See, e.g.*, NINAREH MEHRABI, FRED MORSTATTER, NRIPSUTA SAXENA, KRISTINA LERMAN & ARAM GALSTYAN, A SURVEY ON BIAS AND FAIRNESS IN MACHINE LEARNING 3 (2022), https://arxiv.org/pdf/1908.09635.pdf [https://perma.cc/U7JR-V4QV]; Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, Shmargaret Shmitchell, *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*, *in* FACCT '21: PROCEEDINGS OF THE 2021 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 610 (2021).

112. Germain, *supra* note 104.

113. Kimmo Kärkkäinen & Jungseock Joo, *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation*, *in* 2021 IEEE WINTER CONFERENCE ON APPLICATIONS OF COMPUTER VISION 1547, 1547 (2021).

114. *NIST Study Evaluates Effects of Race, Age, Sex on Face Recognition Software*, NIST (Dec. 19, 2019), https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software [https://perma.cc/C4HJ-LZN2].

115. Christina Swarns, *When Artificial Intelligence Gets It Wrong*, INNOCENCE PROJECT (Sept. 19, 2023), https://innocenceproject.org/when-artificial-intelligence-gets-it-wrong/ [https://perma.cc/GXH4-DFTE].

116. *See, e.g.*, Feng et al., *supra* note 100; Ferguson, *supra* note 25, at 516; Lemley & Casey, *supra* note 27, at 1339.

117. *See* Ayelet Gordon-Tapiero, Paul Ohm & Ashwin Ramaswami, *Fact and Friction: A Case in the Fight Against False News*, 57 U.C. DAVIS L. REV. 171, 180 (2023).

118. Alexandra Ulmer & Anna Tong, *Deepfaking It: America's 2024 Election Collides with AI Boom*, REUTERS (May 30, 2023, 11:17 PM), https://www.reuters.com/world/us/deepfaking-it-americas-2024-election-collides-with-ai-boom-2023-05-30/ [https://perma.cc/3WKW-5JM3].

119. Alex Thompson, *First Look: RNC Slams Biden in AI-Generated Ad*, AXIOS (Apr. 25, 2023), https://www.axios.com/2023/04/25/rnc-slams-biden-re-election-bid-ai-generated-ad [https://perma.cc/A2A5-Z38X].

120. *See, e.g.*, Daniel I. Weiner & Lawrence Norden, *Regulating AI Deepfakes and Synthetic Media in the Political Arena*, BRENNAN CTR. FOR JUST. (Dec. 5, 2023), https://www.brennancenter.org/our-work/research-reports/regulating-ai-deepfakes-and-synthetic-media-political-arena [https://perma.cc/268D-6JFB] ("While the problem long predates the advent of sophisticated AI technology, deepfakes and other synthetic images and audio that mimic election officials or other trusted sources in order to disseminate false information can dramatically amplify voter deception campaigns."); Ben

generate and disseminate false information about a politician or their policies to sway elections.[121] The threat of voter deception further chipping away at whatever remaining faith the public has in the electoral process is dire.[122] There is also fear that those in power may use AI in ways that compromise human rights—for example, the use of AI to enable widespread, continuous spying on the population[123] or automated law enforcement.[124]

So far, these harms have been about human interaction with or use of AI, whether benign or malicious. But concerns about independent system behavior itself have begun to proliferate in the mainstream. In the near term, the widespread adoption of AI in the financial sector may begin with a diversity of options,[125] but will ultimately merge into "technological uniformity" with everyone using the same systems that talk to each other.[126] The fear? That the center does not hold, and systems eventually "run amok, and all end up selling the same thing at the same time, causing a market crash."[127] In the long term, there is the threat of catastrophic AI, such as rogue AI agents.[128] Although this technology does not exist yet, companies

---

Quinn, *Slew of Deepfake Video Adverts of Sunak on Facebook Raises Alarm over AI Risk to Election*, THE GUARDIAN (Jan. 12, 2024, 10:07 AM), https://www.theguardian.com/technology/2024/jan/12/deepfake-video-adverts-sunak-facebook-alarm-ai-risk-election [https://perma.cc/ND2K-QYUN] (finding over one hundred paid ads impersonating the UK prime minister were promoted on social media in one month).

121. *See, e.g.*, JOSH A. GOLDSTEIN ET AL., GENERATIVE LANGUAGE MODELS AND AUTOMATED INFLUENCE OPERATIONS: EMERGING THREATS AND POTENTIAL MITIGATIONS 5 (2023), https://arxiv.org/pdf/2301.04246 [https://perma.cc/2KEZ-4NVN].

122. Adam Edelman, *States Are Lagging in Tackling Political Deepfakes, Leaving Potential Threats Unchecked Heading into 2024*, NBC NEWS (Dec. 16, 2023, 7:00 AM), https://www.nbcnews.com/politics/artificial-intelligence-deepfakes-2024-election-states-rcna129525 [https://perma.cc/8YKY-U4LF].

123. Bruce Schneier, *The Internet Enabled Mass Surveillance. A.I. Will Enable Mass Spying.*, SLATE (Dec. 4, 2023, 11:15 AM), https://slate.com/technology/2023/12/ai-mass-spying-internet-surveillance.html [https://perma.cc/76LZ-6VJ5].

124. Jonathon W. Penney & Bruce Schneier, *A.I. Microdirectives Could Soon Be Used for Law Enforcement*, SLATE (July 17, 2023, 11:25 AM), https://slate.com/technology/2023/07/artificial-intelligence-microdirectives.html [https://perma.cc/AT8E-9UL5] ("Made possible by advances in surveillance, communications technologies, and big-data analytics, microdirectives will be a new and predominant form of law shaped largely by machines. They are 'micro' because they are not impersonal general rules or standards, but tailored to one specific circumstance. And they are 'directives' because they prescribe action or inaction required by law.").

125. Gary Gensler & Lily Bailey, Deep Learning and Financial Stability 4 (Nov. 1, 2020) (unpublished manuscript), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3723132 [https://perma.cc/N6A2-DPUS].

126. *Id.*

127. Austin Weinstein, *With AI Booming, Gary Gensler Wants to Keep Finance Safe for Humans*, BLOOMBERG (Aug. 3, 2023, 4:00 AM), https://www.bloomberg.com/news/articles/2023-08-03/sec-chairman-gary-gensler-discusses-the-risks-to-finance-in-ai [https://perma.cc/M87A-YJTX]; *see also* Gensler & Bailey, *supra* note 125 (manuscript at 24).

128. *See generally* DAN HENDRYCKS, MANTAS MAZEIKA & THOMAS WOODSIDE, AN OVERVIEW OF CATASTROPHIC AI RISKS 5, 34 (2023), https://arxiv.org/pdf/2306.12001 [https://perma.cc/HU25-2AWU].

are actively pursuing it, assuming AI will leapfrog human evolution, and building contingency plans "for when AI wipes away our whole economic system."[129]

Given the breadth and magnitude of AI harms, it is no surprise that top of the list of the White House's research and development budget priorities for 2025 is advancing "trustworthy" AI.[130]

3. Bad Engineering to Harmful AI

At the heart of AI harms is the AI engineering process.[131] The skills and expertise required to build AI, especially complex AI such as deep machine learning models, far outstrip basic software engineering skills.[132] Even a relatively simple AI system entails countless decisions, often without right answers. And every decision relating to *whether* and *how* an AI system is built is another opportunity for error and the introduction of risk. Building safe, secure, and trustworthy AI in the first instance requires engaging in responsible AI engineering practices.

AI engineers impact society directly with every decision, from the decision to use AI to solve a problem in the first place,[133] to the class of AI system they choose to solve the problem.[134] These decisions can be broken out into two categories, or two opportunities for error: technical and ethical.

For some problems, the class of AI system selected can be the line between benefit and harm. An AI class such as knowledge representation, may be excellent at diagnosing esoteric illnesses but hopeless at identifying

129. Steven Levy, *What OpenAI Really Wants*, WIRED (Sept. 5, 2023, 6:00 AM), https://www.wired.com/story/what-openai-really-wants/ [https://perma.cc/A7QC-LZ7Y].

130. Memorandum for the Heads of Executive Departments and Agencies (Aug. 17, 2023) [hereinafter Memorandum], https://www.whitehouse.gov/wp-content/uploads/2023/08/FY2025-OMB-OSTP-RD-Budget-Priorities-Memo.pdf [https://perma.cc/4U78-JKSZ].

131. JESSICA FJELD, NELE ACHTEN, HANNAH HILLIGOSS, ADAM CHRISTOPHER NAGY & MAGHULIKA SRIKUMAR, BERKMAN KLEIN CTR., PRINCIPLED ARTIFICIAL INTELLIGENCE: MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI 56 (2020) (finding that international AI governance principles share a belief "that the behavior of such professionals, perhaps independent of the organizations, systems, and policies that they operate within, may have a direct influence on the ethics and human rights impacts of AI").

132. "[B]uilding for customizability and extensibility of models require teams to not only have software engineering skills but almost always require deep enough knowledge of machine learning to build, evaluate, and tune models from scratch." SALEEMA AMERSHI ET AL., SOFTWARE ENGINEERING FOR MACHINE LEARNING: A CASE STUDY 1–2 (2019) (enumerating unique things about AI that need to be considered in the design phase that are not relevant for traditional software development).

133. ANGELA HORNEMAN, ANDREW MELLINGER & IPEK OZKAYA, CARNEGIE MELLON UNIV., AI ENGINEERING: 11 FOUNDATIONAL PRACTICES 2 (2019).

134. Iqbal H. Sarker, *AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems*, SN COMPUT. SCI., Feb. 10, 2022, at 2.

new treatments,[135] whereas another AI class, such as evolutionary algorithms, may be able to discover a new life-saving drug[136] but lack consistency in diagnosing the common cold. Generally, these decisions are technical ones. There are other problems, such as predictive policing, that AI may not be able to solve at all—or may even worsen.[137] The decision to build these systems is an ethical one. Technical mistakes undermine an AI system's accuracy, whereas ethical mistakes undermine an AI system's legitimacy. AI systems often possess flavors of both.

Technical and ethical mistakes are possible across all varieties of AI. But today, deep machine learning systems abound—and responsible AI engineering is especially critical for this class of AI.[138] Colloquially called "black boxes" for their opaque decision-making processes,[139] these models rely on neural nets, or layers of mathematical structures that are capable of handling unparalleled quantities and varieties of data, with the deeper layers doing the lion's share of the sophisticated data analysis.[140] The more layers, the more information the neural net can process, and the more powerful its predictive capacity.[141] But there is a catch: the more layers there are, the more opaque the decision-making process is, and the less likely a human can understand how the system works, creating more opportunity to introduce risk into the system.

To add to the complexity, machine learning AI models aren't told what to do, but *trained;* it learns on its own *how* to make accurate predictions from data.[142] Instead of ensuring successful predictions by articulating loophole-free instructions, AI engineers improve a machine learning model's accuracy through training and testing. AI engineers surrender control over the model's decision-making process in some ways, which

135. *See* Marta Garnelo & Murray Shanahan, *Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations*, 29 CURRENT OP. BEHAV. SCIS. 17, 17 (2019) (describing symbolic AI as "handcrafted" rather than "learned from data"); STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 23 (4th ed. 2021) (describing early expert systems such as the Mycin system for diagnosing blood infections, which incorporated knowledge "acquired from extensive interviewing of experts").

136. *See generally* Ali Ghaheri, Saeed Shoar, Mohammad Naderan & Sayed Shahabuddin Hoseini, *The Applications of Genetic Algorithms in Medicine*, 30 OMAN MED. J. 406 (2015).

137. *See* Mayson, *supra* note 25, at 2251.

138. *See* Yann LeCun, Yoshua Bengio & Geoffrey Hinton, *Deep Learning*, 521 NATURE 436, 436, 438 (2015).

139. FRANK PASQUALE, THE BLACK BOX SOCIETY 3 (2015).

140. *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).

141. *See* IAN GOODFELLOW, YOSHUA BENGIO & AARON COURVILLE, DEEP LEARNING 438 (2016) ("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.").

142. AMERSHI ET AL., *supra* note 132, at 3 ("During *model training*, the chosen models (using the selected features) are trained and tuned on the clean, collected data and their respective labels.").

makes these other design decisions more important. These intricacies complicate the goal of avoiding mistakes, both technical and ethical.

For example, building a good training dataset is foundational. The model is only as good as its training data: garbage in, garbage out.[143] First, developers must decide whether to procure available data or create their own data. The former reflects real data but can be biased and compromise privacy. Fake, or "synthetic" data, on the other hand, can address data biases[144] and preserve privacy[145] but it can also lead to "model collapse," the implosion of a model's learned behaviors.[146] Avoiding output bias or model collapse is a technical concern, implicating a model's accuracy. Avoiding the use of problematic data, such as datasets compromising privacy or containing known or suspected child sexual abuse material (CSAM),[147] is an ethical concern, implicating a model's legitimacy.

Then, engineers must ensure the training data is "fit" for the model's ultimate use case: does the training data reflect the data the model will process in the real world?[148] Using data from Seattle to train a real estate value prediction model for use in Oaxaca will not yield accurate results—a technical problem. Using training data from 8chan[149] will build a bigoted chatbot—an ethical problem. So, whether found[150] or made,[151] training data

143. *See generally* Brooks Hanson et al., *Garbage In, Garbage Out: Mitigating Risks and Maximizing Benefits of AI in Research*, 623 NATURE 28 (2023).

144. Neil Savage, *Synthetic Data Could Be Better than Real Data*, NATURE (Apr. 27, 2023), https://www.nature.com/articles/d41586-023-01445-8 [https://perma.cc/Y7SV-ZJK9] ("Machine-generated data sets have the potential to improve privacy and representation in artificial intelligence, if researchers can find the right balance between accuracy and fakery.").

145. *Id.*

146. *See* ILIA SHUMAILOV ET AL., THE CURSE OF RECURSION: TRAINING ON GENERATED DATA MAKES MODELS FORGET 1 (2024), https://arxiv.org/pdf/2305.17493 [https://perma.cc/5TDR-6XJG] ("What will happen to GPT-{n} once LLMs contribute much of the language found online? We find that use of model-generated content in training causes irreversible defects in the resulting models, where tails of the original content distribution disappear. . . . Indeed, the value of data collected about genuine human interactions with systems will be increasingly valuable in the presence of content generated by LLMs in data crawled from the Internet.").

147. *See* DAVID THIEL, STAN. INTERNET OBSERVATORY, IDENTIFYING AND ELIMINATING CSAM IN GENERATIVE ML TRAINING DATA AND MODELS, 10–12 (2023), https://purl.stanford.edu/kh752sm9123 [https://perma.cc/DD8J-VB7G].

148. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653, 683–84 (2017).

149. 8chan is an imageboard website that operates with minimal moderation, allowing users to post content anonymously, and has gained notoriety for hosting extremist content, including hate speech, conspiracy theories, and violent content. *See* MAURA CONWAY, RYAN SCRIVENS & LOGAN MACNAIR, INT'L CTR. FOR COUNTER-TERRORISM, RIGHT-WING EXTREMISTS' PERSISTENT ONLINE PRESENCE: HISTORY AND CONTEMPORARY TRENDS 12–14 (2019).

150. *See* Strandburg, *supra* note 24, at 1861 ("[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.").

151. *See* JAMES JORDON ET AL., SYNTHETIC DATA – WHAT, WHY AND HOW? 5 (2022), https://arxiv.org/pdf/2205.03257 [https://perma.cc/8GVW-88QX].

requires analysis, cleaning, and polishing:[152] culling outliers and errors as well as sussing out biases and risks.[153]

Even armed with a robust training dataset, the training process demands expert consideration to avoid harmful errors. First, the engineer must configure the model, setting "hyperparameters," such as the number of layers in a neural net, appropriate for the task at hand.[154] Too many layers, and the model overcomplicates the prediction problem; too few layers and the model oversimplifies it—both are technical errors that undermine real world performance. However, configuration is far from an exact science,[155] which means that without good judgment, an AI engineer can doom a model's chances of success.[156]

After configuration, the model must learn. Sometimes, the learning is *supervised* by checking prediction attempts against an answer key, giving engineers the most control over the training process.[157] Even here, there is ample room to err. And answer keys are not easy to come by; the process of making them is arduous, prone to error, and rife with gray areas—is "crap" profanity for a text classifier and is a taco a sandwich for an image classifier?[158] Mislabeling an answer key can be a technical issue or scrivener's error. However, it can also be an ethical issue, forcing AI engineers to make decisions regarding how to classify the gender of nonconforming individuals for the purposes of an answer key. With

152. KHALID SALAMA, JAREK KAZMIERCZAK & DONNA SCHUT, GOOGLE CLOUD, PRACTITIONERS GUIDE TO MLOPS: A FRAMEWORK FOR CONTINUOUS DELIVERY AND AUTOMATION OF MACHINE LEARNING 6 (2021) ("Data engineering involves ingesting, integrating, curating, and refining data to facilitate a broad spectrum of operational tasks, data analytics tasks, and ML tasks.").

153. *See* AMERSHI ET AL., *supra* note 132, at 3 ("*Data cleaning* involves removing inaccurate or noisy records from the dataset, a common activity to all forms of data science."); *see also* 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 to human experts) to bring the data to a cleaner state suitable for downstream applications and analytics.").

154. *See* GOODFELLOW ET AL., *supra* note 141, at 110 ("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.").

155. *See id.* at 7–8 (explaining that "there is no single correct value for the depth of an architecture").

156. *See id.* at 427 ("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.").

157. *See* Steven A. Israel, Philip Sallee, Franklin Tanner, Jonathan Goldstein & Shane Zabel, *Applied Machine Learning Strategies*, 39 IEEE POTENTIALS 38, 38 (2020) ("The most prominent ML methods in use today are supervised, meaning they require ground-truth labeling of the data on which they are trained.").

158. *See* AMERSHI ET AL., *supra* note 132, at 3 ("*Data labeling* assigns ground truth labels to each record. For example, an engineer might have a set of images on hand which have not yet been labeled with the objects present in the image. . . . Labels can be provided either by engineers themselves, domain experts, or by crowd workers in online crowd-sourcing platforms.").

unsupervised learning, there is no answer key to check prediction accuracy; detecting defects is even more complicated.[159]

Finally, AI engineers must test a model after training before deploying it in the real world, especially because AI regularly performs worse in the real world than diligent testing suggests.[160] This is the final safety net, or opportunity for an engineer to catch an issue before the public is exposed to the model. AI systems must be tested for their ability to serve their stated purpose, to withstand or mitigate risks accurately and reliably, as well as their overall fairness.[161] But testing, like much of AI, is an imprecise science.[162] AI engineers will fail to comprehensively evaluate system accuracy if they do not test for an unforeseen use case—a technical error.

AI engineers must make ethical judgments, prone to error, in assessing performance: how accurate does a system need to be and should it maximize false positives or false negatives? The COMPAS case study is illustrative: a bail-setting and sentencing algorithm was characterized as racist, however the truth is murkier. The algorithm was accurate and racially consistent when evaluating its ability to predict recidivism in assigning risk scores; high scores correlated to high odds of reoffending, regardless of race.[163] However, when the system erred, it erred in a discriminatory way; black defendants who did not reoffend were more than twice as likely to have received a higher risk score. As researchers showed, it is impossible to design a system that is racially consistent in its positive predictability of

159. *See* Percy Liang & Dan Klein, *Analyzing the Errors of Unsupervised Learning*, *in* PROCEEDINGS OF ACL-08: HLT 879, 879–80 (2008).

160. *See* NAT'L ACADS., TESTING, EVALUATING, AND ASSESSING ARTIFICIAL INTELLIGENCE – ENABLED SYSTEMS UNDER OPERATIONAL CONDITIONS FOR THE DEPARTMENT OF THE AIR FORCE 7 (2023) (noting that often "performance of the deployed system in the operational domain was much worse than predicted during the test phase"). *See generally* NAT'L INST. OF STANDARDS & TECH., ARTIFICIAL INTELLIGENCE MEASUREMENT AND EVALUATION WORKSHOP SUMMARY (2022) (acknowledging challenges in evaluating AI systems and noting that performance in controlled environments often does not translate directly to real-world applications).

161. *See* Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz & Andrew D. Selbst, *The Fallacy of AI Functionality*, *in* FACCT '22: PROCEEDINGS OF THE 2022 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 959, 962 (describing four ways that AI systems can fail to function: impossible tasks, engineering failures, post-deployment failures, and communication failures).

162. *See* J. Bergstra, D. Yamins & D.D. Cox, *Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures*, *in* 28 PROCEEDINGS OF THE 30TH INTERNATIONAL CONFERENCE 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").

163. *See* Sam Corbett-Davies, Emma Pierson, Avi Feller & Sharad Goel, *A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased Against Blacks. It's Actually Not That Clear.*, WASH. POST (Oct. 17, 2016, 5:00 AM), https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/ [https://perma.cc/7TEX-HN3P].

recidivism and racially consistent in its error distribution. Deciding which to prioritize is an ethical decision that AI engineers are making today.

The public's interactions with AI, whether positive or negative, are dictated by the myriad decisions made by AI engineers, who may or may not be qualified to be making those decisions.

4. Addressing the Source of the Problem

Given the immense impact AI engineering decisions have on the potential for AI to cause harm,[164] ensuring that AI engineers engage in responsible practices is of paramount importance. Their decisions, technical and ethical, can get baked into the AI system itself, infeasible for users down the line to see, let alone alter or undo.[165] So, AI harms must be avoided or addressed *by design*.[166] AI cannot practically unlearn bad behavior,[167] and backend band-aids or guardrails to solve problems are insufficient[168] or, sometimes, wholly ineffective.[169]

---

164. ROBYN CAPLAN, JOAN DONOVAN, LAUREN HANSON & JEANNA MATTHEWS, DATA & SOC., ALGORITHMIC ACCOUNTABILITY: A PRIMER 22–23 (2018) ("Algorithms are products that involve human and machine learning. While algorithms stand in for calculations and processing that no human could do on their own, ultimately humans are the arbiters of the inputs, design of the system, and outcomes. Importantly, the final decisions to put an algorithmic system on the market belongs to the technology's designers and company."); William D. Smart, Cindy M. Grimm & Woodrow Hartzog, *An Education Theory of Fault for Autonomous Systems*, 2 NOTRE DAME J. ON EMERGING TECHS. 33, 36 (2021) ("In other words, while it is hard to exert meaningful "control" over automated systems to get them to act predictably, developers and procurers have great control over how much they *test* and *articulate the limits* of an automated technology to all the other relevant parties."); Christian Kästner, *Responsible AI Engineering*, MEDIUM (Jan. 7, 2022), https://ckaestne.medium.com/responsible-ai-engineering-c97e44e6c57a [https://perma.cc/NK7L-BS6N] ("Simple implementation decisions like (1) tweaking a loss function in a model, (2) how to collect or clean data, or (3) how to present results to users can have profound effects on how the system impacts users and the environment at large.").

165. *See* EDWARD HU ET AL., LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS 1 (2022), https://arxiv.org/pdf/2106.09685 [https://perma.cc/3W9L-ZTMV] (noting that many design choices are locked into preexisting models and retraining cannot feasibly undo them).

166. *See generally Secure by Design*, CYBERSECURITY & INFRASTRUCTURE SEC. AGENCY, https://www.cisa.gov/securebydesign [https://perma.cc/9QYG-2GKT]; *Privacy by Design*, INTERSOFT CONSULTING, https://gdpr-info.eu/issues/privacy-by-design/ [https://perma.cc/2FLQ-9395]; EUR. COMM'N, ETHICS BY DESIGN AND ETHICS OF USE APPROACHES FOR ARTIFICIAL INTELLIGENCE (2021); John Perrino, *Using 'Safety by Design' to Address Online Harms*, BROOKINGS (July 26, 2022), https://www.brookings.edu/articles/using-safety-by-design-to-address-online-harms/ [https://perma.cc/3XWP-KDJW].

167. Matthew Duffin, *Machine Unlearning: The Critical Art of Teaching AI to Forget*, VENTUREBEAT (Aug. 12, 2023, 8:20 AM), https://venturebeat.com/ai/machine-unlearning-the-critical-art-of-teaching-ai-to-forget/ [https://perma.cc/R8PF-LKRX].

168. Cade Metz, *Researchers Say Guardrails Built Around A.I. Systems Are Not So Sturdy*, N.Y. TIMES (Oct. 19, 2023), https://www.nytimes.com/2023/10/19/technology/guardrails-artificial-intelligence-open-source.html [https://perma.cc/3ZMZ-LC8V].

169. *See* Mayson, *supra* note 25, at 2238 ("[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.").

At a minimum, responsible AI creation requires those building or customizing AI to have the necessary technical and ethical skills to do so responsibly.[170] Recognizing this, higher education programs specifically focus on AI engineering, with several courses (like statistics) in common,[171] and many jobs require specialized training.[172] Unfortunately, AI engineers today "significantly underestimate" the expertise their teams need "nine out of ten times."[173] For example, only one third of developers knows how to properly test systems for risks.[174]

Technical qualification is not enough; societal harms cannot be avoided unless AI engineers consider the public interest in discretionary engineering decisions, such as choosing the metrics for success,[175] but they are not well-equipped to do so. Sometimes, model accuracy must be sacrificed at the

170. *See* SALAMA ET AL., *supra* note 152, at 5 ("This is where ML engineering can be essential. ML engineering is at the center of building ML-enabled systems, which concerns the development and operationalizing of production-grade ML systems. ML engineering provides a superset of the discipline of software engineering that handles the unique complexities of the practical applications of ML."); HORNEMAN ET AL., *supra* note 133.

171. *See, e.g.*, *Artificial Intelligence (AI)*, UC BERKELEY ELEC. ENG'G & COMPUT. SCI., https://www2.eecs.berkeley.edu/Research/Areas/AI [https://perma.cc/Y9A4-SXZ9]; *Schedule: CIS5200 Machine Learning*, CIS 5200 MACH. LEARNING, https://machine-learning-upenn.github.io/calendar/ [https://perma.cc/GP5F-6JCJ] (UPenn); *Artificial Intelligence*, UCLA SAMUELI MASTER ENG'G, https://www.meng.ucla.edu/artificial-intelligence-2/ [https://perma.cc/4FTP-MEKH]; *Doctoral Programs in Computer Science and Engineering*, U.C. SAN DIEGO JACOBS SCH. ENG'G (Aug. 2024), https://cse.ucsd.edu/graduate/doctoral-programs-computer-science-and-engineering [https://perma.cc/DRG5-WFSA]; *Doctor of Philosophy with a Major in Machine Learning*, GA. TECH., https://catalog.gatech.edu/programs/machine-learning-phd/#requirementstext [https://perma.cc/Q2ST-WRG5]; *PhD Course List*, B.U. FAC. COMPUTING & DATA SCIS., https://www.bu.edu/cds-faculty/programs-admissions/phd-degree/phd-course-list/ [https://perma.cc/7LFK-AQYX]; *Ph.D. Degree Requirements*, N.Y.U. COURANT COMPUT. SCI., https://cs.nyu.edu/home/phd/degree_requirements.html [https://perma.cc/7RXW-YXXU]; *PhD in Machine Learning*, CARNEGIE MELLON UNIV., https://www.ml.cmu.edu/current-students/phd-curriculum.html [https://perma.cc/VFZ3-VLWS]; *Ph.D. Requirements*, CORNELL UNIV., https://www.cs.cornell.edu/phd/current-students/phd-requirements [https://perma.cc/KJQ9-SWY5]; *Intelligent Systems, PhD*, UNIV. OF PITT., https://catalog.upp.pitt.edu/preview_program.php?catoid=224&poid=70386&returnto=23022 [https://perma.cc/H3BA-MSPM].

172. Chip Cutter, *The $900,000 AI Job Is Here*, WALL ST. J. (Aug. 14, 2023, 11:46 AM), https://www.wsj.com/articles/artificial-intelligence-jobs-pay-netflix-walmart-230fc3cb [https://perma.cc/KA4L-ZH4U] ("Many of the roles require a degree or advanced experience in computer science, mathematics or data science. Since large language models make judgements based on probability, a strong understanding of statistics is helpful, employers and recruiters say.").

173. HORNEMAN ET AL., *supra* note 133, at 3.

174. Orna Raz, Sam Ackerman & Marcel Zalmanovici, *Managing the Risk in AI: Spotting the "Unknown Unknowns,"* IBM (June 6, 2021), https://research.ibm.com/blog/ai-unknown-unknowns [https://perma.cc/U45J-VNDS].

175. Jan Gogoll et al., *Ethics in the Software Development Process: From Codes of Conduct to Ethical Deliberation*, 34 PHIL. & TECH. 1085, 1087 (2021) ("The developers, experts in this very technical domain, develop the product within the given parameters. Naturally, these parameters will never be completely determined. Therefore, the development team has some leeway in the development of the product.").

altar of the public interest,[176] but figuring out *when* is harder than it seems. Many ethical questions require some degree of training in subject matter beyond technical AI engineering; how, for example, should an AI engineer clean a dataset that contains racial slurs, if filtering out those slurs may also suppress the discourse of marginalized populations seeking to reclaim those terms?[177] We almost certainly do not want engineers making these decisions in isolation, but equipping them to engage outside experts during the development process requires its own kind of training in cross-disciplinary collaboration. Additionally, "[t]he nature of the values puts them inevitably in tension with each other" when building AI systems and there is no objective external "ranking of values" to reference.[178] It may not be possible to achieve the same caliber of accuracy without compromising some amount of user privacy—for which use cases is that trade-off worth it and how are engineers meant to engage in that deliberative process about ethical choices? The legal academic tradition prides itself on its pedagogical priority of training students in this deliberative process; existing AI engineering programs, less so.

B. Drivers of the AI Problem

Bad AI engineering decisions create harmful AI systems, but many AI engineers either lack the skills to do better or are expressly told not to. This section explains that the current landscape is only likely to exacerbate the problem because AI companies are incentivized to build more AI instead of safe AI, and consumers lack the information they need to navigate the minefield.

1. The AI Arms Race

Now that the era of widespread consumer-facing AI has begun, it is unlikely that the status quo will produce reliable, responsible AI.[179] In a case of misaligned incentives, it does not pay for the behemoths building AI to care much about building it responsibly. In fact, responsible AI engineering is a cost—one that industry has chosen to shrug off.

176. Aileen Nielsen, *The Too Accurate Algorithm* 4–9 (Ctr. for L. & Econ., Working Paper No. 09/2022, 2022), https://www.research-collection.ethz.ch/bitstream/handle/20.500.11850/572429/CLE_WP_2022_09.pdf [https://perma.cc/QUH6-NLR2] (explaining how Google rolled back the accuracy of a highly profitable feature because it compromised other human values).

177. Lorena O'Neil, *These Women Tried to Warn Us About AI*, ROLLING STONE (Aug. 12, 2023), https://www.rollingstone.com/culture/culture-features/women-warnings-ai-danger-risk-before-chatgpt-1234804367/ [https://perma.cc/7PH4-S857].

178. Gogoll et al., *supra* note 175, at 1095.

179. *See* NICK HANLEY, JASON F. SHOGREN & BEN WHITE, ENVIRONMENTAL ECONOMICS IN THEORY AND PRACTICE 24 (1997).

Once OpenAI kicked off the AI arms race,[180] Big Tech's old guard and new guard wasted no time in joining the fray.[181] The driving incentive was simple: release more AI fast. AI's big players "all had technology in various stages of development,"[182] but they forbore from releasing it to the public due to concerns about "hallucinations"[183] of false answers, bigotry, economic mayhem, and existential doom. After ChatGPT hit the scene, they could no longer afford to hold back; falling behind meant lost profit and less market dominance. Research labs merged with product factories overnight[184] and employees were hired to propel companies in the AI arms race, despite safety expert warnings.[185] All at once, technology deemed too risky for public consumption was suddenly in every home with an internet connection. Even the government wanted a taste.[186]

The threat of these unchecked market dynamics was foreshadowed years ago. Clearview AI and PimEyes "pushed the boundaries of what the public thought was possible" by releasing facial recognition products for use by the government and the public that can link a person's face to a name, social media profiles, and other online photos of them, including ones they would prefer be kept private.[187] At the time, Google had already built, and chosen not to release, the very same technology because it believed that the technology "was too dangerous to make widely available."[188] As with the AI that followed, Clearview AI and PimEyes did not make "a technological breakthrough; it was an ethical one."[189] Now, facial recognition is

180. Karen Weise, Cade Metz, Nico Grant & Mike Isaac, *Inside the A.I. Arms Race that Changed Silicon Valley Forever*, N.Y. TIMES (Sept. 25, 2024), https://www.nytimes.com/2023/12/05/technology/ai-chatgpt-google-meta.html [https://perma.cc/WL2V-SCD7].

181. Matteo Wong, *The Future of AI Is GOMA*, THE ATLANTIC (Oct. 24, 2023), https://www.theatlantic.com/technology/archive/2023/10/big-ai-silicon-valley-dominance/675752/ [https://perma.cc/9HKB-PQ9F] ("Succession is hardly guaranteed, but a post–Big Tech world might not herald actual competition so much as a Silicon Valley dominated by another slate of fantastically large and powerful companies, some old and some new. Big Tech has wormed it[s] way into every corner of our lives; now Big AI could be about to do the same.").

182. Weise et al., *supra* note 180.

183. *See, e.g.*, Schiffer & Newton, *supra* note 93.

184. Grant, *supra* note 63.

185. *Id.*

186. Memorandum, *supra* note 130, at 2 (calling for the harnessing of AI "to accelerate the Nation's progress"); John D. McKinnon, *Chuck Schumer Joins Crowd Clamoring for AI Regulations*, WALL ST. J. (June 21, 2023, 12:26 PM), https://www.wsj.com/articles/chuck-schumer-joins-crowd-clamoring-for-ai-regulations-7fd8a882 [https://perma.cc/66KE-PYXQ] ("The first issue we must tackle is encouraging, not stifling, innovation.").

187. Kashmir Hill, *The Technology Facebook and Google Didn't Dare Release*, N.Y. TIMES (Sept. 11, 2023), https://www.nytimes.com/2023/09/09/technology/google-facebook-facial-recognition.html [https://perma.cc/CEF9-KHG7].

188. *Id.*

189. *Id.*

everywhere.[190] Most concerningly, the founder of Clearview AI had no real AI engineering talent: he used open-source—code that is freely and publicly available for any use—and read research papers to figure out how to build the product.[191] The lessons learned from this example are twofold: (1) the public is only as safe as the AI ecosystem's weakest link; and (2) it does not take much to be the next company to break the next ethical barrier.

This trend is visible across the industry, with companies responding to macroeconomic pressures by cutting trust, safety, and ethics teams—the teams tasked with ensuring AI is built responsibly.[192] The tech industry has a history of shortchanging trust, safety, and ethics when they threaten the bottom line[193] or company reputation.[194] Even when ethics teams are retained, their input is deprioritized in a software product launch environment, their contributions are harder to quantify and thus underappreciated, and they take on great personal risk when raising ethics issues, especially when they come from marginalized backgrounds.[195]

Big Tech aside, the AI arms race has released the tools to create harmful AI into the public domain, recreating the conditions that gave rise to ubiquitous facial recognition.[196] Although there is a lot to be said for the

190. *See, e.g.*, James Clayton & Ben Derico, *Clearview AI Used Nearly 1m Times by US Police, It Tells the BBC*, BBC (Mar. 27, 2023), https://www.bbc.com/news/technology-65057011 [https://perma.cc/NWV2-88ZC] ("Facial recognition firm Clearview has run nearly a million searches for US police, its founder has told the BBC."); Kashmir Hill, *Facial Recognition Goes to War*, N.Y. TIMES (Apr. 7, 2022), https://www.nytimes.com/2022/04/07/technology/facial-recognition-ukraine-clearview.html [https://perma.cc/L6WR-KJ9P] (explaining the use of Clearview AI's technology, and its risks, in the war on Ukraine); Adlan Jackson, *A Facial-Recognition Tour of New York*, NEW YORKER (Jan. 15, 2024), https://www.newyorker.com/magazine/2024/01/22/a-facial-recognition-tour-of-new-york [https://perma.cc/3V6A-U3VE] (identifying Macy's and Madison Square Garden as users of facial recognition technology).

191. Hill, *supra* note 187.

192. Hayden Field & Jonathan Vanian, *Tech Layoffs Ravage the Teams that Fight Online Misinformation and Hate Speech*, CNBC (May 27, 2023, 7:02 AM), https://www.cnbc.com/2023/05/26/tech-companies-are-laying-off-their-ethics-and-safety-teams-.html [https://perma.cc/W3GU-E3E4] ("The slashing of teams tasked with trust and safety and AI ethics is a sign of how far companies are willing to go to meet Wall Street demands for efficiency . . . .").

193. *See* Billy Perrigo, *How Facebook Forced a Reckoning by Shutting Down the Team that Put People Ahead of Profits*, TIME (Oct. 7, 2021, 11:35 AM), https://time.com/6104899/facebook-reckoning-frances-haugen/ [https://perma.cc/5QAE-67Z2] (detailing information from whistleblower Frances Haugen's testimony that Meta ignored evidence that Instagram makes one-third of women feel worse about their bodies, that proven measures to reduce misinformation were rejected out of hand, and that Meta spends almost all of its budget for keeping the platform safe on English-language content).

194. *See* O'Neil, *supra* note 177.

195. Sanna J. Ali, Angèle Christin, Andrew Smart & Riitta Katila, *Walking the Walk of AI Ethics: Organizational Challenges and the Individualization of Risk Among Ethics Entrepreneurs*, *in* FACCT '23: PROCEEDINGS OF THE 2023 ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 217, 218 (2023).

196. David Evan Harris, Opinion, *AI Is Already Causing Unintended Harm. What Happens when It Falls into the Wrong Hands?*, THE GUARDIAN (June 16, 2023, 3:00 PM), https://www.theguardian.com/commentisfree/2023/jun/16/ai-new-laws-powerful-open-source-tools-meta [https://

value of open-source technologies,[197] especially for AI,[198] there are many pronounced risks, which range from the ability to strip guardrails off of existing powerful models to make uncensored versions[199] to reconfiguring models with the purpose of creating novel biological weapons.[200] Although Meta released its open-source model under the pretext of "democratizing access to . . . AI,"[201] there is reason to believe it was a calculated business decision.[202] The availability of open-source AI also enables small businesses, scrambling to keep up in the AI race, to cut corners and reuse available tools for different purposes without being able to afford the expertise necessary to build off them responsibly.[203] So, a "win" for a few AI companies may have handed the next Clearview AI the tools they need for an "ethical breakthrough."

For their part, AI companies have expressed an interest in self-regulating responsibly,[204] including by making voluntary commitments to safe AI

perma.cc/7P5E-74PV] ("Meta's semi-open source LLaMA and its descendent large language models (LLMs), however, can be run by anyone with sufficient computer hardware to support them – the latest offspring can be used on commercially available laptops. This gives anyone – from unscrupulous political consultancies to Vladimir Putin's well-resourced GRU intelligence agency – freedom to run the AI without any safety systems in place.").

197. Chinmayi Sharma, *Tragedy of the Digital Commons*, 101 N.C. L. REV. 1129, 1140 (2023) ("Open source's most obvious value is that it makes innovative, useful software available to anyone for free.").

198. Jessica Billingsley, *Beyond Corporate AI: Why We Need an Open-Source Revolution*, ROLLING STONE (Nov. 1, 2023), https://www.rollingstone.com/culture-council/articles/beyond-corporate-ai-why-we-need-open-source-revolution-1234867419/ [https://perma.cc/EFN2-JTM5] ("Open-source AI development offers three primary benefits: transparency, community-driven ethics and a counter to monopolistic control. As AI technology grows in complexity, open source provides a transparent foundation for ethical and security standards.").

199. David Evan Harris, *How to Regulate Unsecured "Open-Source" AI: No Exemptions*, TECH POL'Y.PRESS (Dec. 4, 2023), https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/ [https://perma.cc/Q2RZ-B8CV].

200. ELIZABETH SEGER ET AL., CTR. FOR THE GOVERNANCE OF AI, OPEN-SOURCING HIGHLY CAPABLE FOUNDATION MODELS 13–14 (2023), https://cdn.governance.ai/Open-Sourcing_Highly_Capable_Foundation_Models_2023_GovAI.pdf [https://perma.cc/5SUU-CZ6C].

201. *Meta and Microsoft Introduce the Next Generation of Llama*, META (July 18, 2023), https://about.fb.com/news/2023/07/llama-2/ [https://perma.cc/UR5N-SAAD].

202. *See* Harris, *supra* note 196 (explaining that open sourcing a model, such as Meta's LLaMA2, invites the entire community of researchers and independent coders to improve or expand it).

203. CAPLAN ET AL., *supra* note 164, at 16 ("Algorithms are expensive and difficult to build from scratch. Hiring computer scientists, finding training data, specifying the algorithm's features, testing, refining, and deploying a custom algorithm all cost time and money. Therefore, there is a temptation to take an algorithm that already exists and either modify it or use it for something it wasn't designed to do. However, accountability and ethics are context specific. Standards that were set and ethical issues that were dealt with in an algorithm's original context may be problems in a new application.").

204. For example, some companies have released acceptable use policies. These, however, are largely unenforceable, especially for open-source models. *See, e.g.*, *Artificial Intelligence Acceptable Use Policy*, SALESFORCE (Oct. 23, 2024), https://sfdc.co/AI-Policy [https://perma.cc/JFB7-H4B7]; *see also* 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/R3N4-Z7EX] (reporting that membership grew to over one hundred partner organizations spanning thirteen countries and four continents).

engineering[205] and investment[206] practices. But the cat is already out of the bag—harmful AI is already out there and these commitments make no mention of recalling problematic systems.[207] Meta actually released LLaMA2, triggering all the risks of open source AI, just days before promising the White House and the public to engage in responsible AI engineering.[208] Also, the commitments provide no objective standard against which the companies will be held; without real substance,[209] companies are free to, and in fact, incentivized to, hold themselves to a lower standard. This suggests a lot of "virtue signalling" and "it is certainly justified to assume [the commitments] are mere window-dressing."[210] Finally, many of these self-regulation efforts, including the White House summit, are coordinated by a "rarefied club of companies," excluding the countless smaller AI companies from the process.[211]

What's good for the AI industry is not always good for the public. Indeed, the AI arms race demonstrates that when companies are incentivized to win in the market, responsible AI engineering falls by the wayside. However, when these products cause harm, whether by causing an autonomous vehicle to crash or creating synthetic CSAM, society bears the cost, not the companies. A recent survey found that seventy percent of Americans and ninety-two percent of tech experts believe industry needs to invest more in measures to protect the public.[212] This constitutes a market externality; the public values responsible AI engineering more than

205. Press Release, White House, Fact Sheet: Biden-Harris Administration Secures Voluntary Commitments from Leading Artificial Intelligence Companies to Manage the Risks Posed by AI (July 21, 2023), https://www.whitehouse.gov/briefing-room/statements-releases/2023/07/21/fact-sheet-biden-harris-administration-secures-voluntary-commitments-from-leading-artificial-intelligence-companies-to-manage-the-risks-posed-by-ai/ [https://perma.cc/Z8L9-9YGR].

206. The new guidelines "include pledges to 'secure organizational buy-in' within startups around responsible AI, 'forecast AI risks and benefits,' and 'audit and test to ensure product safety.'" Bloomberg, *Top VC Firms Pledge Responsible Strategy on AI Startups*, INV. NEWS (Nov. 14, 2023), https://www.investmentnews.com/fintech/news/top-vc-firms-pledge-responsible-strategy-on-ai-startups-245679 [https://perma.cc/W65E-Z3HJ] (quoting Responsible Innovation Labs).

207. *See* Sue Halpern, *Will Biden's Meetings with A.I. Companies Make Any Difference?*, NEW YORKER (July 24, 2023), https://www.newyorker.com/news/daily-comment/will-bidens-meetings-with-ai-companies-make-any-difference?mc_cid=9b3f171604&mc_eid=f720a42bfb [https://perma.cc/EJN6-Q785].

208. *Id.*

209. *Id.* ("It is also not clear who those experts will be, how they will be chosen, whether the same experts will be tasked with examining all the systems, and by what measure they will determine risk.").

210. *AI Ethics Guidelines Global Inventory*, ALGORITHM WATCH (Apr. 9, 2019), https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/ [https://perma.cc/RTH5-H8KG]; *see also* FJELD ET AL., *supra* note 131.

211. Mohar Chatterjee, *AI Companies Try to Self-Regulate*, POLITICO (Aug. 2, 2023, 4:56 PM), https://www.politico.com/newsletters/digital-future-daily/2023/08/02/ai-companies-try-to-self-regulate-00109502 [https://perma.cc/VDQ4-YLAU].

212. Press Release, MITRE, MITRE-Harris Poll Finds Lack of Trust Among Americans in AI Technology (Feb. 9, 2023), https://www.mitre.org/news-insights/news-release/mitre-harris-poll-finds-lack-trust-among-americans-ai-technology [https://perma.cc/9EB6-TALJ].

companies.[213] As the discrepancy increases, companies are less and less likely to invest appropriately in responsible AI engineering.[214] The problem will not fix itself—we are in an AI safety race to the bottom.[215]

2. Asymmetric Information

When market forces fail to incentivize socially desirable behavior, consumers can shift the balance by exerting pressure through their purchasing power. However, in the AI ecosystem, asymmetric information, or an information imbalance, deprives consumers the information they need to make informed choices.[216] To demand better, the public needs to know they are being harmed and know they can expect better.

Just as no one can look at a pill they were prescribed and evaluate its effectiveness and side effects, neither can they look at an AI system and determine its accuracy and risks.[217] That is why, after a popular medication killed over one hundred people, the FDA took choice over medications out of consumers' hands.[218] While some harmful AI is obvious, other forms of AI snake oil are more pernicious, causing harm through inaccuracy,[219]

---

213. ROBERT COOTER & THOMAS ULEN, LAW & ECONOMICS 44–47 (Denise Clinton, Victoria Warneck, Catherine Bernstock & Lisa P. Flanagan eds., 4th ed. 2004); *see* Michael J. Trebilcock & Edward M. Iacobucci, *Privatization and Accountability*, 116 HARV. L. REV. 1422, 1431–35 (2003). *See generally* A.C. PIGOU, THE ECONOMICS OF WELFARE (1920) (addressing fundamental economic concepts).

214. *See* Mollie Lee, Note, *Environmental Economics: A Market Failure Approach to the Commerce Clause*, 116 YALE L.J. 456, 480 (2006).

215. David Evan Harris, *The Race to the Bottom on AI Safety Must Stop*, CIGI (June 16, 2023), https://www.cigionline.org/articles/the-race-to-the-bottom-on-ai-safety-must-stop/ [https://perma.cc/U248-WLW3].

216. *See* HANLEY ET AL., *supra* note 179, at 68–75; Yafit Lev-Aretz & Katherine J. Strandburg, *Regulation and Innovation: Approaching Market Failure from Both Sides*, 38 YALE J. ON REGUL. BULL. 1, 9–12 (2020) (discussing how information asymmetries in the AI ecosystem can lead to market failures by depriving consumers of the information necessary to make informed choices).

217. *See* CAROLYN COX & SUSAN FOSTER, FTC, THE COSTS AND BENEFITS OF OCCUPATIONAL REGULATION 5–6 (1990) ("One potential source of market failure in professional markets is asymmetric information on quality. Such a failure may occur if it is more difficult for consumers than for the sellers to determine the quality of a service offered. In the extreme case, the quality of a service may not be evaluated even after a purchase.").

218. PETER TEMIN, TAKING YOUR MEDICINE: DRUG REGULATION IN THE UNITED STATES 4, 42, 45, 53–54 (1980) (explaining that once, "most drugs were chosen personally by consumers" and the sale of unsafe drugs was not prohibited until one drug killed over one hundred people and the FDA prohibited the sale of unsafe drugs entirely, concluding that "[c]onsumers . . . were not competent to make their own drug choices").

219. *See* ARVIND NARAYANAN, PRINCETON UNIV. CTR. FOR INFO. TECH. POL'Y, HOW TO RECOGNIZE AI SNAKE OIL, https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf [https://perma.cc/8DY2-33T6]; Sara Collins, *21st Century Snake Oil: The Consequences of Unregulated, Unproven AI*, TECH POL'Y.PRESS (Oct. 13, 2021), https://www.techpolicy.press/21st-century-snake-oil-the-consequences-of-unregulated-unproven-ai/ [https://perma.cc/N7GG-WEXV] ("I wanted to focus on this example, not because of the obvious civil rights harms or how this technology reinforces problematic policing practices, but because independent researchers ultimately found that the algorithm employed by the Chicago Police Department didn't even work.").

insecurity,[220] unsafe practices,[221] or subversive discrimination,[222] and remain at large today. Schools have wrongly punished students for cheating[223] and doctors have made life-or-death diagnoses based on models that are far less accurate than reported.[224] If administrators and physicians, experts in their respective fields, can be fooled by the functionality of AI systems, the average consumer is even more vulnerable.

This issue is exacerbated by the fact that AI harms, however serious, can go undetected indefinitely. In many contexts, AI is a tool intended to help humans make decisions by "find[ing] patterns that are beyond human recognition, often making it difficult to distinguish errors from success and rendering harm from AI errors functionally unforeseeable."[225] A study found that a highly accurate model used to predict death from pneumonia relied on medically nonsensical assumptions in its decision-making—the only reason the researchers knew was because they designed the system to be somewhat transparent.[226] Most AI systems are not.[227] Relying on these seemingly innocuous results would have had life-threatening consequences.

220. Schiffer & Newton, *supra* note 93.

221. *See* Perez, *supra* note 86.

222. Drew Harwell, *A Face-Scanning Algorithm Increasingly Decides Whether You Deserve the Job*, WASH. POST (Nov. 6, 2019, 12:21 PM), https://www.washingtonpost.com/technology/2019/10/22/ai-hiring-face-scanning-algorithm-increasingly-decides-whether-you-deserve-job/ [https://perma.cc/D2CH-GQVX] ("'It's pseudoscience. It's a license to discriminate,' she added. 'And the people whose lives and opportunities are literally being shaped by these systems don't have any chance to weigh in.'").

223. Benj Edwards, *OpenAI Confirms that AI Writing Detectors Don't Work*, ARS TECHNICA (Sept. 8, 2023, 10:42 AM), https://arstechnica.com/information-technology/2023/09/openai-admits-that-ai-writing-detectors-dont-work/ [https://perma.cc/HP5N-QYY6].

224. Rajiv Leventhal, *Sparking Debate, Researchers Question Validity of Epic's Sepsis Prediction Model*, HEA!THCARE INNOVATION (June 24, 2021), https://www.hcinnovationgroup.com/clinical-it/clinical-decision-support/article/21228162/sparking-debate-researchers-question-validity-of-epics-sepsis-prediction-model [https://perma.cc/ME56-GT34] (reporting that a model intended to diagnose sepsis, a condition that causes one in three hospital deaths, predicts cases of sepsis far less than the company building the system officially reports).

225. Selbst, *supra* note 24, at 1321.

226. *Id.* at 1340 (e.g., an AI model predicting pneumonia-related deaths identified asthma as a protective factor based solely on patient outcomes, when in reality, asthma patients had lower mortality rates because they were more likely to receive proactive treatment for their preexisting condition, illustrating that the model, though accurate, relied on flawed and nonsensical assumptions).

227. Cynthia Rudin & Joanna Radin, *Why Are We Using Black Box Models in AI When We Don't Need To? A Lesson from an Explainable AI Competition*, HARV. DATA SCI. REV., Fall 2019, at 1, 3–4 ("Most machine learning models, however, are not designed with interpretability constraints; they are just designed to be accurate predictors on a static dataset that may or may not represent how the model would be used in practice.").

AI companies could,[228] but choose not to,[229] educate their downstream customers and end users about the limitations of the systems they build. Instead, they engage in "AI washing"[230] or "safety-washing"[231]—the use of bold normative commitments to engender public trust in their products.

Even if consumers realized AI is not all that, they would be hard pressed to demand better from AI companies without an understanding of how the technology works and a sense of whether companies could be doing better. Due to "intense corporate secrecy, the contextual nature of AI, and the speed of AI development," the public is often left in the dark regarding how AI causes harms.[232] As a preliminary matter, many AI systems, even simple linear regression models, are complex statistical machines that are beyond the average consumer's ability to comprehend. So, even if consumers noticed they were being harmed, they may not be able to attribute that harm to the use of AI, which means AI shirks the blame. On top of that, AI represents valuable intellectual property, prompting companies to withhold effective AI engineering practices, denying competitors information that could help them build safer systems and denying the public the information needed to prove AI companies could have prevented harms.

Finally, related to the information asymmetries, there is a free rider problem regarding responsible AI. Although in theory, the public craves safer, more secure, and trustworthy AI, *individuals* are not sufficiently incentivized to demand more. Without information needed to measure harms, the average person is likely to leave it to somebody else to do something about harmful AI. Ultimately, this means nobody does anything.[233]

228. Smart et al., *supra* note 164, at 52 (identifying a failure to warn users about the limitations of a system as an education-failure-point, explaining that an example of a proper warning "would be a developer explaining to a procurer, 'you asked us to find people in red shirts, but we do not have a "person in a red shirt detector," we can simply find red-colored, people-shaped blobs—and it will fail if there is a blue light or an apple in range of the sensor'").

229. Katharine Miller, *Introducing the Foundation Model Transparency Index*, STANFORD UNIV. HUMAN-CENTERED A.I. (Oct. 18, 2023), https://hai.stanford.edu/news/introducing-foundation-model-transparency-index [https://perma.cc/L2YP-FNKR] (finding that AI models today fail across most transparency metrics).

230. David A. Shargel, Rachel B. Goldman & Patrick J. Morley, *Compliance Risk After SEC Warning Against 'AI Washing*,*'* LEXOLOGY (Jan. 4, 2024), https://www.lexology.com/library/detail.aspx?g=814b921f-23d2-43a3-a7ab-84a3014caa7a [https://perma.cc/T7NL-TLHF] ("AI washing may occur when a company misleads investors regarding its true AI capabilities.").

231. Seth Lazar & Alondra Nelson, *AI Safety on Whose Terms?*, 381 SCIENCE 138, 138 (2023) ("Sociotechnical approaches recognize and reject 'safety-washing'—giving lip service to safe AI systems, without requisite commitments and practices to ensure this is the case—and call for transparency and accountability to keep companies honest.").

232. Selbst, *supra* note 24, at 1322.

233. COX & FOSTER, *supra* note 217, at 17–18.

C. Roadblocks to Solving the AI Problem

Recognizing there is a problem is the first step, but traditional efforts to fix the problem face an uphill battle. This section explains why, due to the nature of the technology and the outsized influence of the industry, familiar strategies of litigation and regulation will not get the job done on their own.

1. Barriers to Effective Litigation

Although many scholars have explored judicial challenges to harmful AI,[234] private litigation is limited in its ability to hold AI companies accountable for harmful systems. The requirements for criminal and civil law would keep all but the most extreme cases out of court, undermining the goals of punishing bad actors, compensating victims, and deterring future irresponsible behavior.

Criminal law is a poor tool for holding AI companies accountable for harmful systems because to prove a crime, you must also prove *mens rea*, or the defendant's guilty mind.[235] AI systems are generally not self-executing; without the ability to carry out a crime themselves, they are most likely to be accused of assisting in a crime, such as promoting terrorist content. To convict someone of aiding or abetting, or facilitating, a third party's criminal activity, the prosecutor would need to establish the AI company had the *specific intent* or specific purpose of helping another commit a crime, or was at least consciously promoting criminal content.[236] Lesser *mens rea* such as recklessness do not suffice. Given the manner in which AI is built, it can be hard to pinpoint the exact moment in the engineering lifecycle when that *mens rea* was formed, especially in the case of deep machine learning. Even if the AI engineer subjectively possessed the requisite *mens rea*, they cannot act directly on it. They can guide the learning process, but they cannot force the AI system to behave in any specific way.

---

234. *See, e.g.*, Jessica S. Allain, *From* Jeopardy! *to Jaundice: The Medical Liability Implications of Dr. Watson and Other Artificial Intelligence Systems*, 73 LA. L. REV. 1049, 1052 (2013); Claudia E. Haupt & Mason Marks, *AI-Generated Medical Advice—GPT and Beyond*, 329 JAMA 1349, 1350 (2023); Mark F. Grady, *Why Are People Negligent? Technology, Nondurable Precautions, and the Medical Malpractice Explosion*, 82 NW. U. L. REV. 293, 293–94 (1988); Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1, 4 (2018); George S. Cole, *Tort Liability for Artificial Intelligence and Expert Systems*, 10 COMPUT./L.J. 127, 130 (1990); Selbst, *supra* note 24, at 1318 ("And just as with any new technology, negligence law will be called on to adapt and respond to the new threat."); Jane Bambauer, *Negligent AI Speech: Some Thoughts About Duty*, 3 J. FREE SPEECH L. 343, 344 (2023); Mihailis E. Diamantis, *Vicarious Liability for AI*, 99 IND. L.J. 317, 319 (2023).

235. Eugene J. Chesney, *The Concept of Mens Rea in the Criminal Law*, 29 J. CRIM. L. & CRIMINOLOGY 627, 627 (1939).

236. Henderson et al., *supra* note 24, at 627–28, 632.

Tort law presents a more promising avenue for holding companies accountable for harmful AI because general negligence is a more forgiving standard. However, even in the realm of negligence law, plaintiffs face many barriers to victory, which undercuts proposals to subject engineers to a customary standard of care without further reform.[237] For one, given the market's information asymmetries, the harms, or the fact that AI caused them, may even go unnoticed until it is too late.[238] This means that valid suits may never be brought.[239] Even when noticed, constitutional standing requirements will keep the vast majority of AI harms out of court.[240] Those requirements demand plaintiffs demonstrate a concrete and particularized harm that the law considers compensable. Many AI harms, such as privacy violations, biased decision-making, or subtle manipulations, may not satisfy that standard.[241] These harms are often diffuse, incremental, and may not manifest as physical or financial damages typically recognized in tort law. While the harms are deteriorating the fabric of society in the aggregate, courts may treat each individual harm as *de minimis*.[242] Consequently, none but the most egregious cases would survive the standing requirement. For suits that satisfy the requirements, AI will still frustrate attempts to establish causation. How does one prove a system gave the wrong answer when there is no objective right answer?[243] How does one prove the system functioned irresponsibly when no one knows exactly how it functions? Finally, merits of the tort suit aside, the long arm of Section 230 of the Communications Decency Act may bar suits successfully given the fact that AI generates new

237. Choi, *supra* note 32, at 603–09.

238. COX & FOSTER, *supra* note 217, at 7 ("One problem with litigation serving as a mechanism to insure quality, however, is that the consumer must be able to evaluate to some extent the quality of the service received.").

239. MICHAEL J. SAKS & STEPHAN LANDSMAN, CLOSING DEATH'S DOOR: LEGAL INNOVATIONS TO END THE EPIDEMIC OF HEALTHCARE HARM 61 (2021) ("First, the victim must recognize that a harm has occurred and define it as a damaging change from normal or expected (naming). Next, the victim must attribute the cause of the harm to some person or other entity (blaming). And finally, the person must make a decision to confront the harm- doer and demand recompense (claiming). Fewer and fewer victims of tortious injury reach each succeeding stage.").

240. *See, e.g.*, Spokeo, Inc. v. Robins, 578 U.S. 330, 341 (2016); Transunion LLC v. Ramirez, 594 U.S. 413, 442 (2021).

241. SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM 54 (2019) ("These developments are all the more dangerous because they cannot be reduced to known harms—monopoly, privacy—and therefore do not easily yield to known forms of combat."); *see also* Neil Richards & Woodrow Hartzog, *A Duty of Loyalty for Privacy Law*, 99 WASH. U. L. REV. 961, 985 (2021) ("Negligence has also failed to handle privacy issues well because of its intense focus on harm rather than relationships.").

242. Richards & Hartzog, *supra* note 241, at 985.

243. Selbst, *supra* note 24, at 1338 (explaining that AI assists in decision making in circumstances when "the very idea of a measurable truth may not even exist," frustrating attempts to prove the AI system malfunctioned in any way).

content *based on* third-party information it ingested.[244] This is especially true for systems designed to regurgitate memorized information or to attribute outputs to specific third-party sources.

2. Barriers to Effective Regulation

Although far from a foregone conclusion,[245] the vast majority of the public craves government intervention on AI.[246] They are in good company; members of Congress,[247] the White House,[248] and agencies[249] agree that government intervention is necessary to ensure responsible AI. While substantive government regulations must play a role in ensuring that new technologies benefit the public, they cannot be the whole solution. They face substantial barriers to their effectiveness including the rapid pace of innovation, scarcity of experts, and regulatory capture.

Regulation is famously, and by design,[250] unable to update at the pace the science is advancing.[251] While regulations can encourage general categories of responsible behavior, such as promoting security and bias mitigation, they will not be able to *prescribe* and *update* specific strategies

244. Henderson et al., *supra* note 24, at 622 n.112; *see, e.g.*, O'Kroley v. Fastcase, Inc., 831 F.3d 352, 355 (6th Cir. 2016) (finding that Google could not be held liable "for merely providing access to, and reproducing, the allegedly defamatory text" as links and snippets in search engine results).

245. Christiaan Hetzner, *Former Google CEO Eric Schmidt Tells Government to Leave A.I. Regulation to Big Tech*, FORTUNE (May 15, 2023, 10:59 AM), https://fortune.com/2023/05/15/former-google-ceo-eric-schmidt-tells-government-to-leave-regulation-of-ai-to-big-tech-openai-chatgpt-bardai-midjourney/ [https://perma.cc/ZSW6-E6MZ].

246. Press Release, MITRE, *supra* note 212 (reporting that over 80% of the public and over 90% of tech experts support government regulation of AI).

247. *See* McKinnon, *supra* note 186.

248. Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023).

249. Press Release, SEC, SEC Proposes New Requirements to Address Risks to Investors from Conflicts of Interest Associated with the Use of Predictive Data Analytics by Broker-Dealers and Investment Advisers (July 26, 2023), https://www.sec.gov/news/press-release/2023-140 [https://perma.cc/4FTF-TQ3S]; David Garr, *Comments Sought on Amending Regulation to Include Deliberately Deceptive Artificial Intelligence in Campaign Ads*, FED. ELECTION COMM'N (Aug. 16, 2023), https://www.fec.gov/updates/comments-sought-on-amending-regulation-to-include-deliberately-deceptive-artificial-intelligence-in-campaign-ads/ [https://perma.cc/5B9M-BF87]; Press Release, FTC, FTC Staff Report Details Key Takeaways from AI and Creative Fields Panel Discussion (Dec. 18, 2023), https://www.ftc.gov/news-events/news/press-releases/2023/12/ftc-staff-report-details-key-takeaways-ai-creative-fields-panel-discussion [https://perma.cc/4QD2-589C].

250. *Cf.* DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 19 (1980) (describing the conundrum arising from the fact that, by the time technology is in wide enough use to understand the impact, it's often too late to fully control).

251. FJELD ET AL., *supra* note 131, at 57 (quoting the French AI strategy as saying "professionals play an especially important part in emerging technologies since laws and norms cannot keep pace with code and cannot solve for every negative effect that the underlying technology may bring about"); Anna Butenko & Pierre Larouche, *Regulation for Innovativeness or Regulation of Innovation?*, 7 L. INNOVATION & TECH. 52, 66 (2015) ("The 'pacing problem' commonly refers to the situation when technology develops faster than the corresponding regulation, the latter hopelessly falling behind. The metaphor of 'the hare and the tortoise' is often conjured up.").

to accomplish these goals.[252] Where once, cleaning data was the gold standard for addressing biased AI, it is no longer considered enough.[253] This proves problematic, especially given the EU and the US have already hardcoded specifics into AI regulations, such as limiting their mandates to systems that meet a threshold of computational power, or "compute." This seems to be based on the nebulous assumption that compute power is the best proxy for harmful model capabilities. However, not all powerful models meet these compute thresholds. Indeed, at the time of adoption, none of the widely available models did, but many of them were already causing harm. Non-machine learning models may also be powerful—a well configured knowledge representation model can be used to build a bioweapon—but often do not require as much compute power as machine learning models. Additionally, as the technology advances, companies are likely to develop models capable of achieving the same or better capabilities using less compute.[254] In fact, regulation itself can be a powerful driver for new technological advancements by companies hoping to avoid regulation.[255]

Crafting effective regulation also requires some degree of familiarity with the technology being governed.[256] The scarcity of AI engineers in the industry generally,[257] coupled with the government's inability to retain technical talent specifically,[258] means the government is unlikely to build the deep bench of expertise it needs. With AI compensation packages

252. Selbst & Barocas, *Unfair Artificial Intelligence*, *supra* note 32, at 1050 (explaining government's inability to update regulations quickly).

253. Feng et al., *supra* note 100 (explaining that, cleaning data, which was once considered good practice, may not be enough to mitigate the harm of political bias and offering alternative scientifically supported approaches).

254. Helen Toner & Timothy Fist, *Regulating the AI Frontier: Design Choices and Constraints*, CSET (Oct. 26, 2023), https://cset.georgetown.edu/article/regulating-the-ai-frontier-design-choices-and-constraints/ [https://perma.cc/JWG9-M2BU].

255. Casey Newton, *The Case for a Little AI Regulation*, PLATFORMER (Nov. 2, 2023), https://www.platformer.news/p/the-case-for-a-little-ai-regulation [https://perma.cc/962R-R252].

256. *See* Bruce Schneier & Davi Ottenheimer, *Robots Are Already Killing People*, THE ATLANTIC (Sept. 6, 2023), https://www.theatlantic.com/technology/archive/2023/09/robot-safety-standards-regulation-human-fatalities/675231/ [https://perma.cc/DL66-W5Y3] ("Yes, accidents happen. But the lessons of aviation and workplace safety demonstrate that accidents are preventable when they are openly discussed and subjected to proper expert scrutiny.").

257. Cutter, *supra* note 172; Michael Chui, Mena Issler, Roger Roberts & Lareina Yee, *McKinsey Technology Trends Outlook 2023*, MCKINSEY DIGIT. (July 20, 2023), https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-top-trends-in-tech-2023#new-and-notable [https://perma.cc/NG8A-J2KG] (finding that "of 3.5 million job postings in these tech trends[, we] found that many of the skills in greatest demand have less than half as many qualified practitioners per posting as the global average").

258. Natalie Alms, *The People Problem Behind the Government's AI Ambitions*, NEXTGOV/FCW (Nov. 21, 2023), https://www.nextgov.com/artificial-intelligence/2023/11/people-problem-behind-governments-ai-ambitions/392212/ [https://perma.cc/AX2G-G5TH] ("As agencies move to fulfill requirements laid out in Biden's AI executive order, workforce gaps remain 'one of the biggest barriers' according to a White House official.").

circling high six figures,[259] the government simply can't compete.[260] Given the rapid pace of innovation, experts the government is able to hire and retain "will soon have large knowledge gaps without continual updates."[261] This expertise gap may lead to AI regulations that are either infeasible[262] or inadvisable.[263] It may also lead to impotent regulation of software overall.[264] The government seeks to bridge this knowledge gap by consulting with outside experts.[265] But sporadic meetings with an evasive industry is insufficient to build the level of expertise required for effective regulation.[266] Relying on "insular groups of experts they already know,

259. Cutter, *supra* note 172.

260. *See, e.g.*, Maya Kornberg, *Congress is Woefully Unprepared to Regulate Tech*, SLATE (June 25, 2023, 9:00 AM), https://slate.com/technology/2023/06/congress-tech-committee-ai-regulation.html [https://perma.cc/QNL3-BAA5] ("House Science, Space, and Technology Committee staff declined by nearly 45 percent between 1994 and 2016. Hiring caps and salary caps, insufficient funding, and a hiring pipeline that does not proactively seek out scientific and technical expertise are all to blame.").

261. Van Loo, *supra* note 31, at 406.

262. For example, watermarks on AI generated content are required by the White House executive order. Exec. Order No. 14,110, 88 Fed. Reg. 75191 (Oct. 30, 2023). But it is infeasible to unalterably watermark AI. *See* Kate Knibbs, *Researchers Tested AI Watermarks—and Broke All of Them*, WIRED (Oct. 3, 2023, 6:00 AM), https://www.wired.com/story/artificial-intelligence-watermarking-issues/ [https://perma.cc/HJ3Y-QNEA] (explaining that scientific research found that it is functionally impossible to build a watermark for AI generated content that is resilient to removal or alteration attempts).

263. For example, the White House Executive Order 14,110 requires companies building powerful models to red-team their AI systems, but this method is untested. Borrowed from the security world, where red-teaming meant to test the internal security of a system or network, in the AI context, red-teaming refers to the process of testing an AI system from the user standpoint to see how it might break. For example, AI systems may break when it is overloaded with the volume of prompts, when a user inputs programming code instead of conversational text as a prompt, or when a user tricks the AI system into providing prohibited content, such as a manual for how to build a bomb. Jacob Metcalf & Ranjit Singh, *Scaling Up Mischief: Red-Teaming AI and Distributing Governance*, HARV. DATA SCI. REV., Dec. 13, 2023, at 1–2 ("The White House has highlighted AI red-teaming as a central pillar of AI safety in its landmark voluntary agreement with AI developers, and in sponsoring the generative AI red-teaming event at the DEF CON hacker conference in Las Vegas in 2023. Yet this method is largely untested, and it is unclear that it can accomplish the outcomes that policymakers believe it can. We suggest that there are major methodological questions to be addressed before investing so much certainty in this governance strategy.").

264. *See* Nathan Cortez, *Regulating Disruptive Innovation*, 29 BERKELEY TECH. L.J. 175, 192 (2014) (describing the FDA's approach to medical device software as "the archetype of regulatory minimalism"); FDA, PROPOSED REGULATORY FRAMEWORK FOR MODIFICATIONS TO ARTIFICIAL INTELLIGENCE/MACHINE LEARNING (AI/ML) - BASED SOFTWARE AS A MEDICAL DEVICE (SAMD) 3 (2019), https://www.fda.gov/media/122535/download [https://perma.cc/24L3-VKFC] ("The traditional paradigm of medical device regulation was not designed for adaptive AI/ML technologies . . . ."); Framework for Automated Driving System Safety, 85 Fed. Reg. 78058, 78059 (proposed Dec. 3, 2020) (to be codified at 49 C.F.R. pt. 571) (declaring that promulgation of safety standards for automated software systems is "premature" because the development process is "complex and iterative").

265. *But cf.* Kanishka Singh, *US Senate's Schumer: AI Regulations Needed but Should Not Be Rushed*, REUTERS (Sept. 13, 2023, 1:38 PM), https://www.reuters.com/technology/us-senates-schumer-ai-regulations-needed-should-not-be-rushed-2023-09-13/ [https://perma.cc/74FW-UEF7] (warning that "[i]f you go too fast, you can ruin things").

266. Van Loo, *supra* note 31, at 406 ("Complexity, secrecy, and innovation mean that inspectors rely on industry representatives to explain the technology at a facility." (internal quotations omitted)).

many of whom live and work in D.C.,"[267] or going directly to industry[268] risks both underexposure to diverse perspectives on the issue and, worse still, regulatory capture.[269]

The risk of regulatory capture around AI is ever present and more of a threat today than before—AI is the new golden child of the tech industry.[270] Today, many regulators would rather underdeter than chill innovation.[271] Indeed, the predominant approach to AI regulation today is risk regulation, which assumes the AI ecosystem "need only be tweaked at the edges."[272]

Big tech's market power gives it the ability to edge out academia and lock up AI expertise "within a small number of well-funded private companies," further solidifying its dominance.[273] AI's resource demands have led to a brain drain from academia, which cannot compete on resources, to industry.[274] In 2011, about half as many AI PhDs went to industry as they did academia; in 2021, more than double the number of PhDs went to industry.[275] This means that today, industry, not academia, is at the cutting edge of AI.[276] With current estimates pricing machine learning training at $4 million per system today and up to $500 million in less than ten years, the barriers of entry to this industry are *high*,[277] meaning there are unlikely to be many unicorn startups cropping up to challenge the incumbents. This creates a feedback loop, ensuring "the continued accrual of resources and political capital to these companies."[278]

---

267. Kornberg, *supra* note 260.

268. *See* Merlin Stein & Connor Dunlop, *Safe Before Sale*, ADA LOVELACE INST. (Dec. 14, 2023), https://www.adalovelaceinstitute.org/report/safe-before-sale/ [https://perma.cc/N4U6-QP6K] ("Regulatory agencies like the FDA sometimes need industry expertise, especially in novel areas where clear benchmarks have not yet been developed and knowledge is concentrated in industry. Foundation models present a similar challenge. This could raise concerns around regulatory capture and conflicts of interest.").

269. *See, e.g.*, Ben Schreckinger, *The Billionaire Bucks Shaping AI Policy*, POLITICO (Dec. 18, 2023, 4:27 PM), https://www.politico.com/newsletters/digital-future-daily/2023/12/18/whos-really-driving-ai-policy-00132306 [https://perma.cc/89Q3-JUKV] (reporting the RAND corporation, "which has cultivated ties to a growing influence network backed by tech money," has its fingerprints all over the White House Executive Order on AI, with employees concerned about the influence tech money is having on the organization).

270. AMBA KAK & SARAH MYERS WEST, AI NOW, 2023 LANDSCAPE: CONFRONTING TECH POWER 4 (2023) (explaining that we see AI as "synonymous with progress").

271. McKinnon, *supra* note 186 ("The first issue we must tackle is encouraging, not stifling, innovation.").

272. Kaminski, *supra* note 31, at 3.

273. Sam Sabin, *1 Big Thing: Throwing the World's Cyber Arsenal at AI Models*, AXIOS CODEBOOK (Aug. 11, 2023), https://www.axios.com/newsletters/axios-codebook-daad2e2f-da2a-464e-9862-b0e57cc0b67e.htm.l [https://perma.cc/C24L-FRCE].

274. STANFORD UNIV. HUMAN-CENTERED A.I., *supra* note 85, at 245.

275. *Id.*

276. *Id.* at 50 ("In 2022, there were 32 significant industry-produced machine learning systems compared to just three produced by academia.").

277. Duffin, *supra* note 167.

278. KAK & MYERS WEST, *supra* note 270, at 6.

Because AI is "foundationally reliant on resources that are owned and controlled by only a handful of Big Tech firms," these firms have an outsized voice in regulatory conversations,[279] to the exclusion of smaller businesses.[280] For example, while the EU was negotiating its AI Act, companies such as OpenAI claimed a "key victory" when they convinced regulators to drop many proposed restrictions.[281] The AI industry often swarms government officials with lobbyists pushing for anticompetitive regulations.[282] The AI industry has come out in favor of a new agency licensing AI systems,[283] while civil society favors mandatory risk audits as well.[284] Odd as this purportedly pro-regulation stance might seem for tech, "[i]t may not be a coincidence that those companies and their partners have been the strongest advocates of A.I. regulation."[285] Indeed, regulation with "[a] high cost of compliance could limit the number of developers" entering the market,[286] which may "further empower industry leaders, overburden small businesses, and undercut regulators' ability to properly enforce the letter and spirit of the law."[287]

Substantive regulation has an important job in the regulation of emerging technology, but it works best when cast in a supporting role. Indeed, its ability to curtail the novel harms of new technology is enhanced when other forms of regulation uncover hidden risks and diagnose root causes first. On their own though, government bodies are too slow, too uninformed, and too vulnerable to industry sycophancy to be the only bastion of hope for AI regulation.

---

279. *Id.*

280. *See* Emilia David, *AI Regulation Is Taking Shape, but Startups Are Being Left Out*, THE VERGE (Aug. 8, 2023, 9:43 AM), https://www.theverge.com/2023/8/8/23820423/ai-startups-regulation-big-tech [https://perma.cc/Y256-8PNT].

281. Derek Robertson, *France's Mistral Takes a Victory Lap*, POLITICO (Dec. 13, 2023, 4:36 PM), https://www.politico.com/newsletters/digital-future-daily/2023/12/13/frances-mistral-takes-a-victory-lap-00131624 [https://perma.cc/BVW7-APBH]; *see also* Schneier & Ottenheimer, *supra* note 256 ("OpenAI, for example, has reportedly fought to 'water down' safety regulations and reduce AI-quality requirements.").

282. *See* Brendan Bordelon, *As States Move on AI, Tech Lobbyists Are Swarming In*, POLITICO (Sept. 8, 2023, 4:23 PM), https://www.politico.com/news/2023/09/08/tech-lobby-state-ai-efforts-00114778 [https://perma.cc/5HDQ-UMR6].

283. Ryan Tracy, *ChatGPT's Sam Altman Warns Congress that AI 'Can Go Quite Wrong*,*'* WALL ST. J. (May 16, 2023, 1:12 PM), https://www.wsj.com/articles/chatgpts-sam-altman-faces-senate-panel-examining-artificial-intelligence-4bb6942a [https://perma.cc/Q9DK-AM3X].

284. ACCOUNTABLE TECH, AI NOW INST. & EPIC, ZERO TRUST AI GOVERNANCE 7 (2023), https://ainowinstitute.org/publication/zero-trust-ai-governance [https://perma.cc/88JD-6AJW].

285. Wu, *supra* note 88.

286. Stein & Dunlop, *supra* note 268; *see also* Wu, *supra* note 88 ("[P]re-emptive regulation can erect barriers to entry for companies interested in breaking into an industry. Established players, with millions of dollars to spend on lawyers and experts, can find ways of abiding by a complex set of new regulations, but smaller start-ups typically don't have the same resources. This fosters monopolization and discourages innovation. . . . It may not be a coincidence that those companies and their partners have been the strongest advocates of A.I. regulation.").

287. ACCOUNTABLE TECH., AI NOW INST. & EPIC, *supra* note 284, at 4.

II. THE PROMISE OF PROFESSIONALIZATION

Luckily, there is a promising alternative to traditional interventions that would ensure AI serves the public interest by minimizing technical and ethical errors: professionalizing AI engineers.[288] Professionalization refers to the process by which people with jobs are organized into formal communities held to a higher standard. This would ensure that only qualified professionals are building AI, and only in sanctioned ways. Professionalization accomplishes this by establishing academic requirements at accredited universities; creating mandatory licenses to "practice" commercial AI engineering; erecting independent organizations that establish and update codes of conduct and technical practice guidelines; imposing penalties, suspensions, or license revocations for failure to comply with codes of conduct and practice guidelines; and applying a customary standard of care, also known as a malpractice standard, to individual engineer decisions in a court of law. It's a familiar playbook used in medicine, law, accounting, architecture, and every other form of engineering, from civil to electrical. At the end of the day, professionalization seeks to imbue a community of experts with a common purpose, a sense of scientific integrity, and a charge to *do no harm*—an AI Hippocratic oath.[289]

This Part makes the affirmative case for professionalizing AI engineering, explaining the benefits it would confer on society, by overcoming the market pressures driving irresponsible AI practices, and on engineers themselves, by breaking the stranglehold the dominant incumbents have on the field. It will also explain that professionalization avoids the roadblocks to traditional interventions, by conscripting engineers to regulate themselves and establishing a new malpractice cause of action.

The traditional constructs of professionalization, from licensing and practice guidelines to internal discipline, work best to reduce the incidence of technical errors. The sociological underpinnings of professionalization, or its capacity to facilitate a ground-up culture change by formalizing the field and granting it the imprimatur of legitimacy, offers the best chance of galvanizing the inherent desire to "do good" motivating many AI engineers into a discipline-wide reorientation toward ethical and socially responsible behavior.

---

288. Chinmayi Sharma, *Setting a Higher Bar: Professionalizing AI Engineering*, LAWFARE (Dec. 12, 2023, 12:00 PM), https://www.lawfaremedia.org/article/setting-a-higher-bar-professionalizing-ai-engineering [https://perma.cc/FRN9-54C3].

289. *See* Hajar, *supra* note 1, at 156–57.

A. Prioritizing the Public Interest

Ultimately, all the different voices in the AI debate want the same thing: safe and socially beneficial AI. However, today, companies are not incentivized to serve the public interest.[290] Professionalizing AI engineers can overcome this market failure by imposing an affirmative duty on AI engineers to protect society's welfare. This section explains how forcing the field of AI engineering to consider the public interest in establishing and upholding technical standards and codes of conduct minimizes the risk of irresponsible engineering practice creating harmful AI.

1. Establishing Social Responsibility

Professionalization would require AI engineers to behave morally,[291] provide high quality services, and put the public interest before selfish considerations.[292] Because clients, and society at large, are especially vulnerable in the context of professional relationships, "the norm of selflessness [must be] more than lip-service."[293] An incompetent doctor "may fail to diagnose a contagious disease, thus contributing to an epidemic," and the collapse of a building or bridge falls on the local community more than the contracting parties.[294] In that vein, decisions AI engineers make have cascading effects on society and so the field must also put the public interest first. The backstop on every decision, from the first decision, should be: *do no harm*. Professionalization seeks to accomplish

290. *See supra* Section I.C.2.

291. *See* Deborah L. Rhode, *Moral Character as a Professional Credential*, 94 YALE L.J. 491, 493 (1985) ("Moral character as a professional credential has an extended historical lineage."); Harold L. Wilensky, *The Professionalization of Everyone?*, 70 AM. J. SOCIO. 137, 140 (1964) ("But the success of the claim to professional status is governed also by the degree to which the practitioners conform to a set of moral norms that characterize the established professions.").

292. Wilensky, *supra* note 291, at 140 ("These norms dictate not only that the practitioner do technically competent, high-quality work, but that he adhere to a service ideal—devotion to the client's interests more than personal or commercial profit should guide decisions when the two are in conflict."); *see also* Eli Wald & Russell G. Pearce, *Being Good Lawyers: A Relational Approach to Law Practice*, 29 GEO. J. LEGAL ETHICS 601 (2016) (advocating for lawyers to engage in practices that consider the well-being of all parties involved, including clients, colleagues, and the broader community, rather than focusing solely on narrow self-interest or profit maximization).

293. Wilensky, *supra* note 291, at 140; *see also* Richards & Hartzog, *supra* note 241, at 987 ("The core idea animating a duty of loyalty is that trusted parties must make their own interests subservient to those made vulnerable through the extension of trust."); *Walking Through a Minefield: Jonathan Zittrain on the Future of the Internet, Then and Now*, LOGIC(S) (Aug. 1, 2018), https://logicm.io/failure/Jonathan-zittrain-on-the-future-of-the-internet/ [https://perma.cc/CZ4Q-2FSF]; Jack M. Balkin & Jonathan Zittrain, *A Grand Bargain to Make Tech Companies Trustworthy*, THE ATLANTIC (Oct. 3, 2016), https://www.theatlantic.com/technology/archive/2016/10/information-fiduciary/502346/ [https://perma.cc/DD49-E86Y] ("Because doctors, lawyers, and accountants know so much about us, and because we have to depend on them, the law requires them to act in good faith—on pain of loss of their license to practice, and a lawsuit by their clients.").

294. COX & FOSTER, *supra* note 217, at 10.

this by establishing institutions and instilling a culture that minimizes to the extent possible technical and ethical errors.

The process of professionalization necessarily entails *earning* the public's trust by guaranteeing a minimum standard of care—technical competency. Professionalization requires finding "a technical basis" for the field, "assert[ing] an exclusive jurisdiction" over governance, "link[ing] both skill and jurisdiction to standards of training," and in doing so "convince[ing] the public that its services are uniquely trustworthy."[295] As a discipline, *responsible* AI engineering demands a skillset inaccessible to most without substantial education, whether formal or informal. Failure to obtain the relevant skills and abide by best practices results in irresponsible AI engineering decisions that ultimately harm the public. Unfortunately, not everyone building AI has the skills required,[296] which means unskilled software developers are working on technology out of their league. Moreover, regardless of training, there are AI engineers cutting corners[297]—propelling development and releasing products, amidst safety expert warnings.[298] To professionalize, the field would need to exclude individuals who lack the necessary skills and prohibit the building of AI systems in irresponsible ways. Doing so would help protect the public from harmful AI, thereby earning its trust.

History reveals how professionalization elevates the quality of service offered to the public. Accountants and doctors were motivated to professionalize by a desire to earn the public's trust—neither succeeded until they convinced the public that they would prioritize society's welfare over their own self-interest.

Over decades, the process of professionalization has realigned accounting standards with the public's interests, introducing improvements that better protect investors from financial ruin. The field first professionalized after the public witnessed the devastating effects of allowing the industry to behave in self-serving ways. Much like AI engineering, accounting began as an informal practice loosely based on common principles.[299] Lax accounting standards that prioritized direct client interests over the wellbeing of the investing public contributed to more than

295. Wilensky, *supra* note 291, at 138.

296. Raz et al., *supra* note 174 (explaining that only one third of developers know how to properly test AI systems).

297. *See* Ariel Conn, *Can AI Remain Safe as Companies Race to Develop It?*, FUTURE LIFE INST. (Aug. 4, 2017), https://futureoflife.org/recent-news/ai-race-avoidance-principle/ [https://perma.cc/8H3Q-G8XG] ("Cutting corners on safety is really just an act of selfishness.").

298. Grant, *supra* note 63 (explaining how Google's "effort to propel itself to the front of the A.I. pack" pushed Google to launch Bard, despite the product not being ready to safely give life, medical, financial or legal advice).

299. GARY JOHN PREVITS & BARBARA DUBIS MERINO, A HISTORY OF ACCOUNTANCY IN THE UNITED STATES 133, 151 (1998).

one financial crisis.[300] After each incident, the accounting field was forced to reckon with their failure to protect the public and pushed to establish more socially conscientious practice standards.

The process of professionalization transformed the field of medicine from a diffuse community of unqualified and irresponsible practitioners to the industry of well-trained professionals we know today. Physicians originally professionalized in response to a tidal wave of litigation propelled by a staggeringly low opinion of the medical field. In the mid-1800s, perhaps one half of Harvard's medical students could not write.[301] Medical degrees were awarded liberally but left physicians functionally untrained and more likely to make mistakes.[302] Educated surgeons dropped instruments and continued surgery without resterilization; others would probe open wounds with bare hands after wiping their nose or brows.[303] The field was invaded by "gross and ignorant pretenders . . . whose whole existence [was] a continued system of mal-practice."[304] In response, the public hit physicians with an "unprecedented malpractice epidemic,"[305] forcing the medical field to raise their standard of care.

Physicians realized they would be unable to earn the public's trust without incontrovertible proof that their field only consisted of qualified, responsible, upstanding physicians. And so the medical field began the "long trek toward excellence and respectability."[306] They began by advocating for mandatory licenses, following countries such as Britain, France, and Germany which had "stringent licensure laws that were designed 'to develop, foster, and advance true scientific medicine.'"[307] Schools lengthened the course of study, introduced basic science, employed full-time professors, instituted challenging entrance exams, and required a wide array of courses, from chemistry to the "young science of

300. George O. May, *Influence of the Depression on the Practice of Accountancy*, 54 J. ACCOUNTANCY 336, 336 (1932) (explaining the impact of the Great Depression on the development of good accounting practices); WALLACE E. OLSON, ACCOUNTING PROFESSION: YEARS OF TRIAL, 1969–1980, at 1 (1982) (explaining how after the 1969 financial crisis, "[p]ractitioners recognized that they had responsibilities to shareholders, credit grantors, and other third parties"); Daniel L. Goelzer, Bd. Member, Pub. Co. Acct. Oversight Bd., Address on Lessons from Enron: The Importance of Proper Accounting Oversight (July 26, 2006) (explaining how after Enron, the passage of the Sarbanes-Oxley Act of 2002 introduced new rules to further ensure that accountants did their work with more than their clients' interests in mind); S.P. Kothari & Rebecca Lester, *The Role of Accounting in the Financial Crisis: Lessons for the Future*, 26 ACCT. HORIZONS 335, 335 (2011) (explaining how accounting standards contributed to the 2008 financial crisis).

301. KENNETH ALLEN DE VILLE, MEDICAL MALPRACTICE IN NINETEENTH-CENTURY AMERICA: ORIGINS AND LEGACY 73–74 (1990).

302. *Id.*

303. *See id.* at 219.

304. *Id.* at 88.

305. *Id.* at 23; *see also id.* at 87.

306. *Id.* at 90.

307. *Id.* at 87.

bacteriology" for graduation.[308] Efforts by professional associations such as the American Medical Association (AMA) "helped standardize and unify medical beliefs and treatments."[309] Collectively, licenses, education, and standards set by professional organizations reclaimed the esteemed status of the physician and elevated patient care.

By professionalizing, the field of AI engineering can follow the accounting and medical professions in reorienting their priorities to serve the public interest, thereby providing a higher quality of service. By imposing an expectation of technical competency in carrying out services,[310] professionalization would deny AI engineers the right to cut corners with their work or work on projects for which they are unqualified. Doing so would risk their livelihood, their ability to practice in the first place.

Professionalization also has an ethical component; it cultivates a culture of social responsibility and moral decision-making. In medicine and law, there are ethical guidelines that forswear dishonesty, prohibit certain forms of profiteering, and demand certain forms of conduct even outside the professional context. The provision of most services requires some form of ethical decision-making, and professionalization seeks to train individuals to engage in those deliberative processes responsibly, through formal education, guidelines, and ethical hotlines, and it seeks to foster an innate desire to be moral.

Other professionalized fields have leveraged the organizing power of professional associations to gather its participants, raise ethical questions, and develop policies to promote moral behavior in the provision of services. The Architectural profession condemned the use of their skills to support oppressive correctional facilities, the medical profession foreswore the ability to use its professionals to carry out executions, and the legal profession punished its own for denying a legitimate election. These formal ethical deliberations can manifest as formal codes of ethical conduct and generate training to promote compliant behavior. For example, the National

308. *Id.* at 198; *see also* Maxwell J. Mehlman, *Professional Power and the Standard of Care in Medicine*, 44 ARIZ. ST. L.J. 1165, 1174 (2012) (explaining how raising the standard for medical education led to the shuttering of diploma mills).

309. *See* DE VILLE, *supra* note 301, at 90; *see also* Mehlman, *supra* note 308, at 1175 ("Through its licensure and educational reform efforts, the medical profession by the early twentieth century had gained effective control not only over entry into the profession, but over the general contours of the standard of care expected of its members."); Mary Anne Bobinski, *Law and Power in Health Care: Challenges to Physician Control*, 67 BUFF. L. REV. 595, 603 (2019).

310. *Artificial Intelligence Engineering*, CARNEGIE MELLON UNIV., https://www.sei.cmu.edu/our-work/artificial-intelligence-engineering/ [https://perma.cc/R46X-4R36] ("The discipline of AI engineering aims to equip practitioners to develop systems across the enterprise-to-edge spectrum, to anticipate requirements in changing operational environments and conditions, and to ensure human needs are translated into understandable, ethical, and thus trustworthy AI.").

Association of Realtors now requires fair housing training as part of its ethics education. Starting in 2025, both new and existing members will need to complete two hours of fair housing training every three years, alongside their required ethics training. In these courses, realtors would be trained in fair housing laws, what constitutes a protected class, discriminatory practices, and advertising and marketing compliance—training which helps real estate professionals identify and guard against more pernicious forms of discrimination through proxies and deleterious stereotypes.[311] The National Association of Emergency Medical Technicians advocates for cultural competency training for its Emergency Medical Technicians (EMTs)—motivated by a desire to provide its services in accordance with moral social behavior. It integrates cultural sensitivity and competency training into its continuing education programs. Courses like Advanced Medical Life Support and Prehospital Trauma Life Support now include components that focus on understanding the cultural context of patient care, highlighting how cultural factors can affect trauma assessment and treatment decisions. The incorporation of ethical standards can enhance the quality of services delivered beyond respecting social expectations of empathetic care for all. Cultural competency initiatives have shown improved communication among Emergency Medical Service personnel and their patients, better patient outcomes, and a reduction in health disparities.[312]

In this way, professionalization would also force engineers to treat ethics "as both a software design consideration and a policy concern."[313] For example, it would demand more deliberation around Google's decision to pay parents fifty dollars to use their children's faces in training data.[314] While on one hand, the program improves Google's AI products and therefore its bottom line, on the other hand, we may question the morality of paying parents to compromise their children's privacy. Similarly, professionalized AI engineers would recognize that, however cheap the

311. NAT'L ASS'N OF REALTORS, CODE OF ETHICS AND STANDARDS OF PRACTICE OF THE NATIONAL ASSOCIATION OF REALTORS (2024), https://www.nar.realtor/sites/default/files/documents/2024-nar-coe-standards-of-practice-2023-12-21.pdf [https://perma.cc/AH7P-24BC]; *Code of Ethics: Code of Ethics Training*, NAT'L ASS'N REALTORS, https://www.nar.realtor/about-nar/governing-documents/code-of-ethics/code-of-ethics-training [https://perma.cc/5J8P-EBLM].

312. *See* NAT'L ASS'N OF EMERGENCY MED. TECHNICIANS, 2021/2022 EDUCATION CATALOG 2, 6–7, 11–12 (2022), https://www.naemt.org/docs/default-source/education-documents/education-catalog-2022-09-08-2021-final.pdf [https://perma.cc/6827-5CN6].

313. *See* HORNEMAN ET AL., *supra* note 133, at 3.

314. *See* Joseph Cox, *Google Contractor Pays Parents $50 to Scan Their Childrens' Faces*, 404 MEDIA (Jan. 4, 2024, 9:23 AM), https://www.404media.co/google-telus-pays-50-to-scan-childrens-eyelid-shape-and-skin-tone/ [https://perma.cc/45XD-XHBN].

training data, using information scraped from certain parts of the internet are virtually certain to create bigoted and misogynistic systems.[315]

Beyond design decisions, a broader duty to society would also require AI engineers to reconsider whether to build an AI system in the first place. Professionalization permits, and sometimes requires, practitioners to deny service on moral grounds. Architects refused to build unethical prisons[316] and the AMA prohibits physicians from assisting in torture.[317] There are some problems AI simply cannot solve without harming the public.[318] For example, scholars have found that hiring algorithms evaluating an applicant's personality are essentially modern-day phrenology wholly divorced from real science.[319] They simply do not work. Being asked to build a bogus system is not nearly enough justification for a true professional to do so.

Even when an AI system works, professionalized AI engineers may still object to its existence. A report from the Government Accountability Office found that the Federal Bureau of Investigation has done tens of thousands of face recognition searches, but only 5% of the 200 agents with access to the technology completed the associated training—training which was proven to decrease the incidence of false identifications.[320] In light of these realities, socially conscious AI engineers may refuse to develop these systems altogether until conditions on the ground change. Prisons, torture, and facial recognition in law enforcement are all state sanctioned—relying on the government to curb abuse will not work. Forcing the engineers building these systems to consider the public's wellbeing is a better bulwark for society.

315. *See, e.g.*, KENT K. CHANG, MACKENZIE CRAMER, SANDEEP SONI, DAVID BAMMAN, SPEAK, MEMORY: AN ARCHAEOLOGY OF BOOKS KNOWN TO CHATGPT/GPT-4, at 9 (2023), https://arxiv.org/pdf/2305.00118 [https://perma.cc/LR44-87RS] (explaining large language models, are more likely to regurgitate verbatim passages that emerge more frequently in their training data).

316. *See* Zuckerman, *supra* note 107.

317. AM. MED. ASSOC., AMA CODE OF MEDICAL ETHICS: 9.7.5 TORTURE, https://code-medical-ethics.ama-assn.org/sites/amacoedb/files/2022-08/9.7.5.pdf [https://perma.cc/PP2Y-JNPS].

318. *See* Zuckerman, *supra* note 107 ("There are real, current problems with AI systems that might be sufficient reason to rethink their deployment.").

319. *See* Ajunwa, *supra* note 25, at 1190–94; *see also* Raji et al., *supra* note 161, at 960–62; Narayanan, *supra* note 219.

320. Khari Johnson, *FBI Agents Are Using Face Recognition Without Proper Training*, WIRED (Sept. 25, 2023, 5:07 PM), https://www.wired.com/story/fbi-agents-face-recognition-without-proper-training/?utm_source=substack&utm_medium=email [https://perma.cc/559D-DEMH]; *see also* Andrew D. Selbst, danah boyd, Sorelle A. Friedler, Suresh Venkatasubramanian & Janet Vertesi, *Fairness and Abstraction in Sociotechnical Systems*, *in* FAT* '19: PROCEEDINGS OF THE CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY 59 (2019) (explaining that sometimes risk regulation fails when parties like the police use technologies like facial recognition in a context for which it was not designed without understanding the risks and error rates).

2. Introducing a Duty to Third Parties

Professionalization establishes this broader duty to society in large part by subjecting AI engineers to third-party liability. Over a tumultuous history, tort law eventually arrived at the conclusion that professionals should be held liable for the harms they cause certain third parties.[321] Originally, the Supreme Court feared the "absurd consequences"[322] of allowing third parties to sue professionals for economic losses—the classic "opening the floodgates of litigation" argument. Eventually, however, developments in negligence law[323] led courts to chip away at the privity requirement in personal injury cases, with blockbuster decisions like *MacPherson v. Buick Motor Co*.[324] Eventually courts moved away from a privity requirement for third parties entirely, allowing third parties to recover from professionals for malpractice.[325] One of the foundational cases establishing this new right acknowledged that subjecting professionals to third party liability for foreseeable economic harms would "elevate the cautionary techniques of the accounting profession."[326] Of note, a professional cannot disclaim malpractice liability to third parties through contract.[327]

Malpractice liability is a potent incentive to be socially responsible. The threat of it alone can generate the very standards that may be lacking in industry today.[328] Scholars have written about tort liability's behavioral effects: the benefit of "internal deterrence." This theory argues that negligence achieves internal deterrence by "articulating and reinforcing, through concrete applications, obligations to act carefully that tend already to be observed" by individuals in the relevant community.[329] In this way,

321. *See* Jay M. Feinman, *Liability of Accountants for Negligent Auditing: Doctrine, Policy, and Ideology*, 31 FLA. ST. U. L. REV. 17, 22 (2003).

322. Savings Bank v. Ward, 100 U.S. 195, 203 (1879).

323. *See* Fleming James, Jr., *Limitations on Liability for Economic Loss Caused by Negligence: A Pragmatic Appraisal*, 25 VAND. L. REV. 43, 47 (1972).

324. 111 N.E. 1050, 1053 (N.Y. 1916).

325. *See* Biakanja v. Irving, 320 P.2d 16, 19 (Cal. 1958); *see also* Rusch Factors, Inc. v. Levin, 284 F. Supp. 85, 91 (D.R.I. 1968) ("Why should an innocent reliant party be forced to carry the weighty burden of an accountant's professional malpractice? Isn't the risk of loss more easily distributed and fairly spread by imposing it on the accounting profession, which can pass the cost of insuring against the risk onto its customers, who can in turn pass the cost onto the entire consuming public? Finally, wouldn't a rule of foreseeability elevate the cautionary techniques of the accounting profession?").

326. *See Rusch Factors*, 284 F. Supp. at 91.

327. *See* Beacon Residential Cmty. Ass'n v. Skidmore, Owings & Merrill LLP, 327 P.3d 850, 861–62 (Cal. 2014).

328. *See* Choi, *supra* note 32, at 622 ("Software developers should be designated as professionals, not as a retrospective reward for exhibiting professional attributes, but as a prospective incentive for software communities to develop and enforce basic professional customs.").

329. *See* John C.P. Goldberg & Benjamin C. Zipursky, *Accidents of the Great Society*, 64 MD. L. REV. 364, 367 (2005).

malpractice law can be shaped by norms for which the profession has already found "a reason and a motivation for heeding" and it can, in turn, "add to this incentive for heeding it, and the disincentive for flouting it."[330]

Today, software developers generally evade liability, and so a new duty to third parties harmed by AI is likely to "have substantial impact."[331] It's a compelling incentive for very cautious behavior—when malpractice is on the table, you don't cut it close; you steer well clear of noncompliance. The EU currently estimates that introducing liability would generate "€54.8 billion in added value for the EU economy by stepping up the level of research and development in AI and in the range of €498.3 billion if other broader impacts, including reductions in accidents, health and environmental impacts and user impacts are also taken into consideration."[332]

There are, however, many barriers to the effective functioning of ex post civil liability. Professionalization's ability to impose a duty to third parties does not live and die by the tort system. It has alternative means of holding its own responsible for the impact they have on society. Namely, through internal disciplining and professional standards. In 1981, the Hyatt Regency Walkway in Kansas City collapsed, in a devastating failure of engineering practices that killed 114 people. In the aftermath, an investigation revealed significant miscommunications between the engineers and fabricators as well as a failure to adhere to proper safety standards. In short, the primary civil engineer behaved laxly, verbally approving an unsafe design change to a part of the walkway he was not managing, assuming the engineer actually in charge of that part would pick up the slack and catch any mistakes before it was built. This, of course, did not happen. The American Society of Civil Engineers Committee on Professional Conduct (CPC) was not persuaded by the primary civil engineer's argument that each engineer in the design process was responsible for his or her own part of the work—so, he may have verbally approved a faulty design, but the ultimate responsibility should have fallen on the engineer actually managing that part of the walkway. The CPC found that the engineer whose seal is used to approve the final design is legally and ethically responsible for all elements of the structural design and, therefore, "vicariously responsible" for all

330. *Id.* at 396; *see also* Theodore Silver, *One Hundred Years of Harmful Error: The Historical Jurisprudence of Medical Malpractice*, 1992 WIS. L. REV. 1193, 1202 ("The evolution of negligence doctrine mirrors the evolution of a culture. Hence, the legal duties that each citizen owes all others are fundamentally a function of social development.").

331. *See* Choi, *supra* note 32, at 626.

332. *See* Tatjana Evas, EUR. PARLIAMENTARY RSCH. SERV., CIVIL LIABILITY REGIME FOR ARTIFICIAL INTELLIGENCE 1 (2020).

components of the walkway, whether they be civil engineering responsibilities or structural engineering responsibilities.[333]

Like construction, AI engineering also involves large, complex teams of employees, each with their own areas of expertise and responsibility, often with minimal communication between groups. This chaotic status quo leads to the same kind of diffusion of responsibility when something goes wrong and resistance to accountability for "someone else's work." A professional association could clarify that the person who gives final approval to an engineering decision for an AI product bears responsibility for all components of the work involved, from data curation to red-teaming. On the back end, a disciplinary tribunal could enforce this notion of responsibility by penalizing the AI engineer whose "seal" of approval allowed an error-ridden AI product to enter the public domain.

There won't be an overnight culture change in a profession that has long had libertarian undertones and an industry that urges its engineers to "[m]ove fast and break things."[334] But, unlike voluntary standards, professionalization can force AI engineers to consider the broader implications of their decisions and hold them to concrete, defensible standards.

B. Empowering AI Engineers

Professionalization does not benefit society at the cost of AI engineers; there's something in it for them too. Professionalization empowers AI engineers to defend their decisions in the court of law and against their employers. This section argues that, when sued, a customary standard of care protects diligent engineers from frivolous or misguided suits. It also asserts that AI engineers can justify going against a company directive to engage in irresponsible behavior or withhold socially beneficial information by citing to professional mandates.

1. Using Customary Care as a Malpractice Shield

Today, the public is largely on tech's side, but it is unlikely to stay that way. When AI's honeymoon period ends, lawsuits will surge as they did in accounting and medicine, and juries will not be kind to members of the

333. Tara Hoke, *A Question of Ethics: The Hyatt Regency Walkway Collapse*, AM. SOC'Y CIV. ENG'RS (Jan. 1, 2007), https://www.asce.org/publications-and-news/civil-engineering-source/civil-engineering-magazine/article/2007/01/the-hyatt-regency-walkway-collapse [https://perma.cc/QNR7-2GGA].

334. *See* Nick Bilton, *Silicon Valley's Most Disturbing Obsession*, VANITY FAIR (Oct. 5, 2016), https://www.vanityfair.com/news/2016/10/silicon-valley-ayn-rand-obsession [https://perma.cc/7GDJ-2ZF6].

industry. However, if AI engineers professionalize, the customary care standard would protect those who are deserving.[335]

Because customary care forces judges and juries to defer to expert opinions on the standard of care,[336] AI engineers are unlikely to suffer liability for meritless claims.[337] Courts hold professionals to the customary care standard instead of ordinary negligence,[338] which means that a decision's reasonableness is assessed by another expert in the field.[339] Where custom, or widespread practice, can be ignored for an ordinary negligence claim, it *must* be respected in a malpractice suit.[340] Indeed, before customary care, juries were prone to ignore even unanimous expert testimony—when experts disagreed, juries had even freer rein to substitute true expertise with their own.[341] With an abstruse discipline like medicine or AI engineering, where some errors are inevitable,[342] juries are especially likely to mistrust expert opinion.[343] In applying customary care, malpractice

335. *Cf.* Richard N. Pearson, *The Role of Custom in Medical Malpractice Cases*, 51 IND. L.J. 528, 545 (1976).

336. *See* Catherine T. Struve, *Doctors, the Adversary System, and Procedural Reform in Medical Liability Litigation*, 72 FORDHAM L. REV. 943, 945 (2004) ("In many malpractice cases, each element of the claim—standard of care, breach, causation, and damages—requires medical expert testimony.").

337. David M. Studdert et al., *Claims, Errors, and Compensation Payments in Medical Malpractice Litigation*, 354 NEW ENG. J. MED. 2024, 2029 (2006) (finding that malpractice suits are quite good at distinguishing valid claims from frivolous ones and that when malpractice makes a mistake, it is most often in the doctor's favor); Philip G. Peters, Jr., *Doctors & Juries*, 105 MICH. L. REV. 1453, 1492 (2007).

338. *See* Mehlman, *supra* note 308, at 1176–88.

339. *See* Hathorn v. Richmond, 48 Vt. 557, 558–59 (1876) ("[T]he question is, how much skill is [the physician] bound to have and to exercise in order that he should not be liable for a disastrous result? It is a little difficult to define it--you can only describe it or illustrate it. The ordinary expression is, "ordinary skill." That means, such skill as doctors in the same general neighborhood, in the same general lines of practice, ordinarily have and exercise in like cases. If a doctor does in a case what the average class of doctors are accustomed to do and would do in such a case, then he exercises what is meant by ordinary skill in a given case. If he exercises such skill, then he is not liable."); *see also* THOMAS M. COOLEY & JOHN LEWIS, 2 A TREATISE ON THE LAW OF TORTS OR THE WRONGS WHICH ARISE INDEPENDENTLY OF CONTRACT., 1392–93 (3d ed. 1906) ("A physician is entitled to have his treatment tested by the rules of the school of medicine to which he belongs and whose system he professes to practice, and he is only bound to exercise such reasonable care and skill as is usually exercised by physicians of the school in good standing.").

340. *See* Kenneth S. Abraham, Essay, *Custom, Noncustomary Practice, and Negligence*, 109 COLUM. L. REV. 1784, 1791 (2009) ("[T]he custom rule does not require that [custom] evidence be 'taken into account,' but only permits the jury to consider custom evidence if it wishes to do so. The jury may wholly disregard custom evidence without violating the custom rule.").

341. *See* DE VILLE, *supra* note 301, at 53–54.

342. *See* Choi, *supra* note 32, at 614–15 (explaining that "bad outcomes are endemic to the practice of [software engineering]" and "are mainly attributable to inherent uncertainties in the science of the profession").

343. *See* Tim Cramm, Arthur J. Hartz & Michael D. Green, *Ascertaining Customary Care in Malpractice Cases: Asking Those Who Know*, 37 WAKE FOREST L. REV. 699, 702–03 (2002); *see also* Choi, *supra* note 32, at 616 (suggesting that when a profession is sufficiently complex and prone to unpreventable errors, juries are especially limited in their ability to discern reasonable from unreasonable behavior).

law all but guarantees that comporting with the edicts of science guards against the long arm of the law.

Given the nascency and constantly evolving nature of AI standards, reasonable AI engineers may hold differing opinions; customary care is especially protective here.[344] Science often lacks bright line rules. While some things are clearly wrong, such as using data collected from Mumbai to predict the weather in La Paz, most decisions at the cutting edge of a field have more than one right answer. A customary care standard holds room for differing opinions, thereby allowing experimentation in a field of science to flourish, until clear winners emerge.[345] It is not malpractice to hold a minority opinion—for example, that all generative AI should be extractive or "quotes only," meaning the system only regurgitates verbatim text from identifiable sources—as long as it is a valid school of thought supported by responsible authority. [346]

If AI engineers uphold the principles of professionalism and maintain the public's trust,[347] then they benefit from the application of a customary care standard.

2. Resisting Big Tech's Dominance

Professionalization, like many grassroots efforts to organize labor, empowers practitioners against their employers.[348] The threat of malpractice serves as a shield against corporate directives to cut corners or otherwise compromise professional standards. Engineers can cite the need to develop industry standards to demand the right to share information outside of the company's walled gardens. Finally, a portable license goes wherever they go—loosening big tech's stranglehold on the AI industry today.

Professionalization can overcome the strong market incentives driving companies to build AI systems irresponsibly. Although company motivations would remain unchanged, engineers would be armed with the ability to refuse to comply with requirements if compliance conflicts with professional standards. For example, a hospital cannot demand a physician

344. *See* Claudia E. Haupt, *Unprofessional Advice*, 19 U. PA. J. CONST. L. 671, 707–08 (2017).

345. *See, e.g.*, Gary T. Schwartz, *The Beginning and the Possible End of the Rise of Modern American Tort Law*, 26 GA. L. REV. 601, 664 (1992).

346. *Cf.* Thomas L. Hafemeister, Leah G. McLaughlin & Jessica Smith, *Parity at a Price: The Emerging Professional Liability of Mental Health Providers*, 50 SAN DIEGO L. REV. 29, 31–32 (2013) (arguing, for example, that higher damages are awarded because science has provided a better understanding of psychological harm).

347. *See* Pearson, *supra* note 335; DE VILLE, *supra* note 301, at 73–74 (describing abandonment of the customary care standard when physicians lost the public's trust through low quality services and the eventual return to customary care once the medical field earned the trust of the public again).

348. KAK & MYERS WEST, *supra* note 270, at 11 ("[W]orker-led organizing . . . has emerged as one of the most effective approaches to challenging and changing tech company practice and policy.").

double the number of patients seen in a day if doing so would force them to compromise patient care, thereby committing malpractice. Similarly, if the AI engineering community concludes that open-sourcing highly capable models runs counter to the public interest and prohibits it, Meta cannot force its employees to commit malpractice.

Professionalization can also stem the tide of irresponsible demands in the first instance, with the added benefit of saving AI engineers the cost of malpractice insurance. Just as hospitals and private practices pay for physician malpractice insurance,[349] AI employers are likely to pay for their engineers. Not only does this shift the burden of insurance onto the employer, but it also incentivizes management to refrain from putting employees in compromising situations, because doing so would expose the company to liability as well. For example, hospitals have established morbidity boards, which review patient deaths to both inform physicians how to improve care and assure the hospital that patients are not dying due to malpractice.[350]

In the same vein, professionalization would encourage AI engineers to collaborate across the industry to set minimum standards for AI systems across different domains. When a profession is forced to establish industry-wide concrete practice guidelines, multi-stakeholder processes would demand open cooperation.[351] Today, AI companies are loath to share information related to negative user incidents with their products. Once, the airline industry was much the same. However, today, airline professional associations specifically encourage information sharing among aviation professionals, airlines, and airplane manufacturers about safety issues and incidents in an effort to issue industry-wide safety bulletins, which can lead to changes in safety protocols or updates to aircraft designs.[352] Armed with a more complete picture of user interactions with AI, individual AI engineers, industry-wide organizations, and even regulators would be better positioned to warn the public of risks and update practice guidelines to mitigate concerns.

349. *Malpractice Insurance: What You Need to Know*, 3 J. ONCOLOGY PRAC. 274, 274 (2007).

350. Juliet Higginson, Rhiannon Walters & Naomi Fulop, *Mortality and Morbidity Meetings: An Untapped Resource for Improving the Governance of Patient Safety?*, 21 BMJ QUALITY & SAFETY 576 (2012).

351. *Cf.* MARKUS ANDERLJUNG ET AL., FRONTIER AI REGULATION: MANAGING EMERGING RISKS TO PUBLIC SAFETY 16–22 (2023), https://arxiv.org/pdf/2307.03718 [https://perma.cc/W9VE-83ZY] (arguing that mechanisms to create and update safety standards for responsible frontier AI development and deployment should be developed via multi-stakeholder processes).

352. INT'L CIV. AVIATION ORG., SAFETY MANAGEMENT MANUAL (SMM) (4th ed. 2018); INT'L AIR TRANSP. ASS'N, IOSA PROGRAM MANUAL 12–14 (13th ed. 2019); *Reporting Safety Issues*, FED. AVIATION ADMIN. (Sept. 15, 2023), https://www.faa.gov/aircraft/safety/report [https://perma.cc/YV9S-X8HB].

Beyond reputational concerns, companies restrict information sharing because of intellectual property and trade secret concerns. However, professionalized fields have drawn a line in the sand: basic public safety is not a competitive advantage. For example, the AMA officially condemns "[t]he intentional withholding of new medical knowledge, skills, and techniques from colleagues for reasons of personal gain" because it "is detrimental to the medical profession and to society."[353] Responding to this unified stance by the profession, a law was passed preventing surgical procedure patent-holders from asserting their rights against other medical professionals. Just as a successful heart disease intervention must be shared with the profession at large—so too must AI companies be required to share information related to serious system risks or promising precautionary practices.[354]

The boon of information sharing goes beyond increasing technical competency; it can promote ethical behavior as well. Today, there are very few legal obligations on civil engineers to be sustainable. However, construction companies can earn a sustainability rating from the US Green Building Council (USGBC) through its Leadership in Energy and Environmental Design (LEED) certification, which evaluates how environmentally friendly and energy-efficient a project is.[355] USGBC is run by industry professionals such as contractors, engineers, and scientists. In a productive feedback loop, clients expressed a desire for more sustainable buildings, which led the profession to strive for higher LEED ratings to attract more conscious clients. While this might seem like pure market forces at play, there is more going on. A traditional economics story would predict that companies withhold strategies or techniques that achieve greater sustainability. On the contrary, USGBC urges professionals to share their methods across companies to improve sustainability for all projects.[356] It even posts "case studies" on certain projects to formalize best practices. The public's ethical priorities with AI have a better chance of being met if profession at large collaborated on strategies to achieve those ends.

353. *AMA* Code of Medical Ethics' *Opinions on Patenting Procedures and Devices*, 12 AM. MED. ASS'N J. ETHICS 96 (2010), https://journalofethics.ama-assn.org/article/ama-code-medical-ethics-opinions-patenting-procedures-and-devices/2010-02 [https://perma.cc/9Y9K-DWJS].

354. *See* David Weitzner, *Push for AI Innovation Can Create Dangerous Products*, THE CONVERSATION (July 19, 2022, 3:53 PM), https://theconversation.com/push-for-ai-innovation-can-create-dangerous-products-186101 [https://perma.cc/ZV4P-SL8J] ("In other words, a company should not stand out for finding ethical ways to run its business. Ethical commitments should be the minimum expectation required of all who compete.").

355. *LEED Rating System*, U.S. GREEN BLDG. COUNCIL, https://www.usgbc.org/leed/certification [https://perma.cc/R29Q-J63A].

356. U.S. GREEN BLDG. COUNCIL, FROM VISION TO ACTION: USGBC ADVANCING BUILDING DECARBONIZATION 2 (2024), https://www.usgbc.org/sites/default/files/2024-03/USGBC_Advancing-Building-Decarbonization_2024.pdf [https://perma.cc/BG44-BDNG].

Finally, licensing the line AI engineer empowers entrepreneurs to go up against the Goliaths of industry. Armed with a license assuring investors and the public of their abilities, they are free to leave their Big Tech employers to start new ventures. Conversely, licensing companies or models forces the available talent to coalesce around big players that society has every reason to believe are not making AI decisions that prioritize the public's interest. Licensed AI engineers would be free to leave the major players, start their own ventures, and boast their verified credentials to garner investors and clients. This benefits not only engineers but also the public, by fostering a more competitive AI landscape.

Professionally-incentivized information sharing could further empower individual AI engineers to unsettle concentrated industry power. In the agricultural industry, the Open Source Seed Initiative (OSSI) is an organization founded by a group of plant breeders, farmers, and advocates for sustainable agriculture who were concerned about the increasing concentration of power in the commercial seed industry. Large companies would develop and patent genetically modified or hybrid seeds, which restricts farmers from saving and replanting seeds from their crops, forcing them to repurchase seeds year over year, further entrenching the commercial seed industry's stranglehold on the market. OSSI responded to this by providing "open-source" seeds, allowing farmers, breeders, and researchers to freely access and develop new plant varieties without the constraints of patents or restrictive licensing agreements.[357] By organizing the professionals in the field and facilitating information sharing, OSSI challenged the dominant market players to the benefit of individual farmers and the public at large. If AI engineers were able to organize in similar fashion and share meaningful research openly, the market would be less reliant on the handful of companies dominating the AI manufacturing space today.

C. Overcoming Roadblocks to Alternative Proposals

If the US government tries to fill the void of technical standards, lack of expertise and regulatory capture may generate inadequate regulations that quell competition. Nor can society rely on traditional legal causes of action to hold AI accountable, given barriers to effective litigation.

357. CR LAWN, OPEN SOURCE SEED INITIATIVE, RESTORING OUR SEED COMMONS: THE NEED FOR CLARITY ABOUT INTELLECTUAL PROPERTY RIGHTS 1–3 (2019), https://osseeds.org/wp-content/uploads/2019/11/OSSI-IP-Lawn-Restoring-Our-Seed-Commons.pdf [https://perma.cc/CQS4-ZNZJ]; Maywa Montenegro de Wit, *Beating the Bounds: How Does 'Open Source' Become a Seed Commons?*, J. PEASANT STUD., 2017, at 4–5.

Professionalization, on the other hand, has better odds of promoting the public interest.

Professionalization conscripts the very experts building AI to the cause of building AI responsibly. As they do with medicine, policymakers would be wise to defer to experts in the field on questions of right and wrong. The government is not in the business of legislating on the appropriate uses of a stent in the artery versus cardiac bypass surgery.[358] Regulation leaves those determinations to the qualified cardiologists and cardiothoracic surgeons.

It behooves society to engage the experts in setting standards for AI engineering because, without them, we have no hope of staying abreast of the science.[359] Standards evolve over time and must continually adapt to novel circumstances.[360] Things that were once true may be disproven.[361] Or, ideas once dismissed may gain traction among experts "long before they are perceived as legitimate by the public."[362] Following the science, medicine eventually discarded bloodletting and embraced some forms of stem cell treatments. However, regulators and the common law are slow to the uptake.

Additionally, AI engineers are better suited to discern good research from bad. Perhaps AI nutrition labels,[363] source-checks for chatbots,[364] well-calibrated uncertainty mechanisms,[365] particular third-party evaluation

358. Transcript of Oral Argument at 31, N.C. State Bd. of Dental Exam'rs v. FTC, 574 U.S. 494 (2015) (No. 13-534) (noting that a state may quite reasonably want a "group of brain surgeons to decide who can practice brain surgery in this State" and not want "a group of bureaucrats deciding that").

359. *See* JOHN DAL PINO, COUNCIL OF AM. STRUCTURAL ENG'RS, DO YOU KNOW THE STANDARD OF CARE? 8–9 (2014), https://docs.acec.org/pub/18803059-a2fd-2d06-cc39-a6d1dd575265 [https://perma.cc/F7JQ-FTDM] (explaining that certain engineering strategies and technologies that are the norm today weren't before).

360. *See* MARCEL BOERSMA ET AL., FLEXIBLE CATEGORIZATION FOR AUDITING USING FORMAL CONCEPT ANALYSIS AND DEMPSTER-SHAFER THEORY (2022), https://arxiv.org/pdf/2210.17330 [https://perma.cc/CXS6-CL8T].

361. *See* Claudia E. Haupt, *Licensing Knowledge*, 72 VAND. L. REV 501, 509 (2019) ("Processes of professionalization bring with them the advent of licensing regimes.").

362. *See id.*

363. *AI Nutrition Facts*, TWILIO, https://nutrition-facts.ai/ [https://perma.cc/MZC9-CWNJ] (including information such as what data it was trained on and whether there is a human in the loop, among other things).

364. *See* Ryan Heath & Ina Fried, *Google's Bard Launches Fact-Check Features*, AXIOS (Sept. 19, 2023), https://www.axios.com/newsletters/axios-ai-plus-08ac9e78-b6a4-4507-a537-7befad0dd983.html [https://perma.cc/GGA3-6LAK] (explaining that Bard has a new feature where it will check line by line if web content supports each of its claims); *see also* Henderson et al., *supra* note 24, at 609–11 (explaining how retrieval-augmented systems provide a chatbot answer to a prompt with links to the sources from which the portions of the content originated).

365. *See generally* Meiqi Sun, Wilson Yan, Pieter Abbeel & Igor Mordatch, *Quantifying Uncertainty in Foundation Models via Ensembles*, *in* WORKSHOP ON ROBUSTNESS IN SEQUENCE MODELING, 36TH CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (2022).

tools,[366] or more user-friendly AI interfaces[367] will be the next research idea that, after a groundswell of support from the engineer community, transforms into accepted standard practice. The field of AI engineers will know first.

Conscripting experts to the job of enforcing defensible standards gives society a better chance of catching when an AI engineer is doing shoddy work. Given the opacity of the information ecosystem and the inscrutable technology itself, consumers and regulators are hard pressed to identify an irresponsibly designed AI system after the fact. In a professionalized world, AI engineers who fail to audit training data for privacy violations or problematic biases can be reported by their colleagues or the public to professional boards that investigate the complaint and issue a penalty where appropriate—the field would police its own members.[368]

Moreover, as with medicine, AI engineering is already specializing into distinct domains and will continue to do so as the field matures. The wealth of literature on AI engineering is expanding exponentially,[369] making it functionally impossible for experts, let alone regulators, to be equally competent in all forms of AI. The right and wrong way to train or fine-tune a specific model requires substantial expertise not just in linear regression or neural networks, but also in the intended application of the model, whether for medical diagnostics or national security.[370] Professionalization channels the expertise that industry already possesses to set a high, defensible bar *for a specific domain*. An added bonus: engineers, famously skeptical of technocrats, are more likely to comply with standards they had a hand in setting.

Opening individual engineers to malpractice claims for irresponsible AI bypasses the barriers raised by tort law's unimaginative conception of harms and liberal application of Section 230 to provide the public with a real

366. *See, e.g.*, LAURA GUSTAFSON ET AL., FACET: FAIRNESS IN COMPUTER VISION EVALUATION BENCHMARK (2023); Anthony M. Barrett et al., *Can We Manage the Risks of General-Purpose AI Systems?*, TECH POL'Y.PRESS (Dec. 5, 2023), https://www.techpolicy.press/can-we-manage-the-risks-of-generalpurpose-ai-systems/ [https://perma.cc/Z2LQ-EHNG]; *The Generative Assessment Project*, ARTHUR, https://www.arthur.ai/gap [https://perma.cc/9PWB-Z628]; *A Holistic Framework for Evaluating Foundation Models.*, CTR. FOR RSCH. ON FOUND. MODELS, https://crfm.stanford.edu/helm/lite/latest/ [https://perma.cc/2H4Q-RR2X]; Emily Dinan et al., *SafetyKit: First Aid for Measuring Safety in Open-Domain Conversational Systems*, *in* 1 PROCEEDINGS OF THE 6OTH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTUTATIONAL LINGUISTICS 4113 (2022).

367. Amelia Wattenberger, *Why Chatbots Are Not the Future*, WATTENBERGER, https://wattenberger.com/thoughts/boo-chatbots [https://perma.cc/J3KR-4TL9].

368. Quinta Jurecic, *The Professional Price of Falsehoods*, KNIGHT FIRST AMEND. INST. COLUM. UNIV. (Mar. 23, 2023), https://knightcolumbia.org/content/the-professional-price-of-falsehoods [https://perma.cc/Z7YY-WSK4].

369. *Growth in AI and Robotics Research Accelerates*, NATURE (Oct. 12, 2022), https://www.nature.com/articles/d41586-022-03210-9 [https://perma.cc/3Y7A-P4P6].

370. *See* CAPLAN ET AL., *supra* note 164, at 22–23.

chance at redress—if not in court, then before a disciplinary tribunal that does not demand the same constitutional requirements. It can also help plaintiffs surmount the challenge of proving causation. If standards demand documentation of decisions,[371] like a physician's requirement to record consultation notes, plaintiffs have a better chance of finding where in the development process something went awry.[372] Malpractice is familiar with and capable of navigating organizational pecking orders to apportion liability, which is especially important for a field like AI engineering that, like medicine and law, involves large teams including nonprofessionals.[373] Moreover, subjecting AI engineers to individual liability is likely to generate a derivative insurance market, ensuring that plaintiffs can recover when they are harmed and are not faced with judgment-proof defendants.

III. PROFESSIONALIZATION IN PRACTICE

Professionalization entails more than a collective agreement to take the job more seriously. This Part uses historical examples to illustrate the necessary conditions for professionalizing a field. It also addresses the primary concern opponents to this approach share—professional protectionism—and offers recommendations to preempt those risks.

A. The Process of Professionalizing AI Engineers

Historically, most fields have traced similar paths to professionalization, beginning with the creation of formal higher education programs coalescing around similar curricular requirements.[374] Then, growing out of these higher education programs, professional associations emerge, setting semiformal standards and codes of conduct for the field.[375] Professionalization does not reach its apex, however, until the field claims exclusive jurisdiction over its

371. *See* NAT'L INST. OF STANDARDS & TECH., ARTIFICIAL INTELLIGENCE RISK MANAGEMENT FRAMEWORK (AI RMF 1.0) (2023) (requiring multiple forms of documentation); *Responsible AI Guidelines*, DEF. INNOVATION UNIT, https://www.diu.mil/responsible-ai-guidelines [https://perma.cc/SPZ6-GRL9]; ACCOUNTABLE TECH, AI NOW INST. & EPIC, *supra* note 284, at 6–7.

372. *Responsible AI Guidelines*, *supra* note 371; ACCOUNTABLE TECH, AI NOW INST. & EPIC, *supra* note 284.

373. Wilensky, *supra* note 291, at 146 ("An increasing percentage of professionals work in complex organizations (scientists, engineers, teachers, architects, even lawyers and physicians)."); Michael D. Scott, *Tort Liability for Vendors of Insecure Software: Has the Time Finally Come?*, 67 MD. L. REV. 425, 475 (2008) ("How does a plaintiff establish that the defects in the software were due to the malpractice of the 'professionals' who worked on the product and not those who would be deemed non-professionals?").

374. Wilensky, *supra* note 291, at 143–44.

375. *Id.* at 144.

governance through mandatory licenses, rigorous evaluation of qualifications, and vigilant internal disciplining.[376]

AI engineering has already made headway in establishing formal higher education programs to train entrants to their field in relevant subject matter and to provide hands-on experience. Many top schools offer AI programs at the undergraduate and graduate level, with many courses in common across them, suggesting a standardization of curriculum is already underway.[377] However, there is substantial room for improvement. Very few AI degree programs today include safety, security, or ethics as mandatory courses in their syllabi.[378] To exercise socially conscious AI engineering practices, AI engineers must be educated about the risks their technology creates. Accordingly, on the path to professionalization, AI degrees should require cross-disciplinary courses that cover the subject matter in the penumbra of the discipline.[379]

Higher education is not enough; AI engineering must also become a licensed profession. As a preliminary matter, without a formal licensing regime contingent on compliance with mandatory standards, courts will refuse to hold AI engineers to a customary standard of care.[380] More

376. *Id.* at 145–46.

377. *See, e.g.*, sources cited *supra* note 171.

378. *See* Christina Pazzanese, *Trailblazing Initiative Marries Ethics, Tech*, HARV. GAZETTE (Oct. 16, 2020), https://news.harvard.edu/gazette/story/2020/10/experts-consider-the-ethical-implications-of-new-technology/ [https://perma.cc/6VTS-RRFA] (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").

379. *Cf.* STEPHEN P. TURNER, LIBERAL DEMOCRACY 3.0: CIVIL SOCIETY IN AN AGE OF EXPERTS 28 (2003) ("Around every core of 'expert' knowledge is a penumbra, a domain in which core competence is helpful but not definitive, in which competent experts may disagree, and disagree because the questions in this domain are not decidable in terms of the core issues that define competence."); THE AM. BD. OF PSYCHIATRY & NEUROLOGY, 2022 GENERAL INFORMATION AND BOARD POLICIES (2021), https://www.abpn.com/wp-content/uploads/2020/07/2021_ABPN_General_Information_and_Board_Policies.pdf [https://perma.cc/9GLH-SSM6] (explaining that psychiatrists must learn neurology and neurologists must learn psychiatry).

380. Hosp. Comput. Sys., Inc. v. Staten Island Hosp., 788 F. Supp. 1351, 1361 (D.N.J. 1992) ("Professionals may be sued for malpractice because the higher standards of care imposed on them by their profession and by state licensing requirements engenders trust in them by clients that is not the norm of the marketplace. When no such higher code of ethics binds a person, such trust is unwarranted."); Superior Edge, Inc. v. Monsanto Co., 44 F. Supp. 3d 890 (D. Minn. 2014) (holding the same); THOMAS H. KOENIG & MICHAEL L. RUSTAD, GLOBAL INFORMATION TECHNOLOGIES: ETHICS AND THE LAW 55 (1st ed. 2018) ("[C]ourts do not recognize computer malpractice because this field does not have a governing body (such as a state bar association), an enforceable code of professional ethics or licensing laws."); RAYMOND T. NIMMER, JEFF C. DODD & LORIN BRENNAN, LAW OF COMPUTER TECHNOLOGY § 9.30 (4th ed., Thomson Reuters 2023) (concluding that although "programming requires significant skill and effective consultation requires substantial business and technical knowledge," it cannot be considered a profession because its practice "is not restricted or regulated at present by state licensing laws" nor is it "substantial self-regulation or standardization of training" by professional association).

importantly, licensing ensures competence.[381] The history of medical professionalization is instructive: a mandatory licensing regime enables the field to weed out the untrained or noncompliant members of their community, by accrediting education programs, rigorously evaluating skill, and disciplining those who fall short of mandatory practice guidelines—thereby raising the minimum quality of service.[382] Standards need teeth—licenses raise the stakes of noncompliance.

Licensing boards are the engines of professional self-governance: they monitor entrants, establish qualifications, dictate practice guidelines, and discipline those who stray from the flock.[383] The board's jurisdiction and makeup is determined by law.[384] Although licensing is usually done by states, the federal government can and should create an AI licensing body for consistency, given the technology's boundaryless reach.[385] A licensing body's organic statute should diversify board membership to include practitioners and lay persons, or members of the public,[386] to represent the public's interests. Technology companies often struggle to align AI systems with the diverse and sometimes conflicting values and priorities of the public, as studies highlight the difficulty of accurately predicting what guardrails and ethical frameworks users expect from these technologies.[387]

---

381. Haupt, *supra* note 361, at 522.

382. ACCREDITATION COUNCIL FOR GRADUATE MED. EDUC., FREQUENTLY ASKED QUESTIONS: PSYCHIATRY (2023), https://www.acgme.org/globalassets/pdfs/faq/400_psychiatry_faqs.pdf [https://perma.cc/J7DK-TP9U].

383. *See generally* Drew Carlson & James N. Thompson, *Policy Forum: The Role of State Medical Boards*, 7 ETHICS J. AM. MED. ASS'N 311 (2005).

384. N.Y. PUB. HEALTH LAW § 230 (McKinney 2023).

385. *See* Gabriel Scheffler, *Unlocking Access to Health Care: A Federalist Approach to Reforming Occupational Licensing*, 29 HEALTH MATRIX 293, 350–52 (2019) (stating that "most scholars who have examined the issue have concluded that Congress possesses the Constitutional authority to preempt state licensing laws . . . ."); Timothy Bonis, *Is a Federal Medical License Constitutional?*, HARV. L. PETRIE-FLOM CTR. (Jan. 3, 2023), https://blog.petrieflom.law.harvard.edu/2023/01/03/is-a-federal-medical-license-constitutional/ [https://perma.cc/7BUU-7GX6]; *see, e.g.*, *Become a Pilot*, FED. AVIATION ADMIN. (July 9, 2024), https://www.faa.gov/pilots/become [https://perma.cc/F7UE-9PK8] (discussing federal pilot licenses); Flytenow, Inc. v. FAA, 808 F.3d 882 (D.C. Cir. 2015) (requiring a web developer of flight simulation software to obtain a federal license); SEC, REGULATION OF INVESTMENT ADVISERS BY THE U.S. SECURITIES AND EXCHANGE COMMISSION (Mar. 2013) (discussing investment advisor federal licenses); *Operator Licensing*, U.S. NUCLEAR REGUL. COMM'N (Oct. 7, 2024), https://www.nrc.gov/reactors/operator-licensing.html [https://perma.cc/D9VE-E437] (discussing federal nuclear operator licenses).

386. *Understanding New York's Medical Conduct Program - Physician Discipline*, N.Y. STATE DEP'T HEALTH (Mar. 2016), https://www.health.ny.gov/publications/1445/ [https://perma.cc/BT8C-3F5J].

387. *Report Overview*, POL.IS, https://pol.is/report/r3rwrinr5udrzwkvxtdkj [https://perma.cc/PC6F-FRNT] (e.g., finding that public opinion on AI governance principles is highly diverse, with distinct opinion groups prioritizing different values; for example, 56% of participants agreed that AI should prioritize marginalized communities, while 25% disagreed, and 18% were uncertain, illustrating the difficulty of achieving consensus on such principles).

A licensing board's first order of business will be to define the scope of the profession: what does it mean to be an AI engineer. As a preliminary matter, it is important to distinguish the practice of AI engineering from the science of AI engineering. Traditionally, licenses are reserved for those who hold themselves out to the public as experts providing a valuable product or service. There is a robust world of academic research and open-source contributions to the field of AI engineering. While regulation over these activities may be warranted, they need not fall under a licensing regime. AI engineers exploring the bounds of this new technology in research labs or diffuse open-source communities are not building systems for public consumption, just as physicians researching new medical interventions in labs are not practicing medicine.[388]

Then, a licensing body must determine the scope of activities that constitutes AI engineering. For example, AI engineering must include training models, but does it include compiling the training data? AI engineering likely includes model configuration, but does it include model testing as well? Rather than create a static definition in law, delegating to a licensing body of expert and lay perspectives to define "AI engineering" allows the definition to be flexible. Today, nurse practitioners are licensed to do work that was traditionally the exclusive purview of physicians, and dental hygienists have taken over tasks originally reserved to dentists—over time, the field may create different tiers or licenses, or allow unlicensed engineers to do work that today is only trusted to the qualified few.

After drawing the boundaries of the discipline, licensing boards and professional associations must establish and enforce standards governing both technical and ethical decisions.[389] While AI codes of ethics have been in vogue, standards are more than just moral commitments. They include concrete, scientifically-backed guidelines around technical, practice decisions.[390] For example, the American Psychiatric Association's Diagnostic and Statistical Manual of Mental Disorders (DSM Manual) is "deemed the bible of psychiatry."[391] It is considered the "authoritative guide to the diagnosis of mental disorders" and contains "descriptions, symptoms,

388. Wilensky, *supra* note 291, at 141 ("The main public for the scientist is fellow-scientists, who are in a position to judge competence; the main public for the professional is clients or employer-clients, who usually cannot judge competence.").

389. Choi, *supra* note 32, at 637 ("Professional principles must flow upward from the practices of the profession.").

390. FJELD ET AL., *supra* note 131, at 59 (finding through a systematic mapping of international AI governance frameworks that many countries believe "those who build and implement AI systems should be guided by established professional values and practices," with some highlighting the scientific method as "the bedrock for technological innovation, including AI").

391. Ralph Slovenko, *The DSM in Litigation and Legislation*, 39 J. AM. ACAD. PSYCHIATRY & L. 6, 6 (2011).

and other criteria" for diagnosis.[392] Its eminent status is justified in part by the fastidious process of creating it. In a massive endeavor that lasted six years, thousands of experts across a range of disciplines conducted literature reviews, proposed preliminary standards, tested them in the field, invited independent perspectives on them, and subjected them to public scrutiny.[393] Since its publication, the DSM Manual has been updated numerous times,[394] welcoming public input each time.[395] Although AI engineering has begun the process of developing standards through National Institute of Standards and Technology (NIST)[396] and the Association for Computing Machinery Code of Ethics,[397] these standards are threadbare in comparison to the DSM Manual and other medical practice guidelines. To raise the floor of AI engineering, standards must be prescriptive, meticulous, and subject to public scrutiny.

Some deny that AI has scientific standards today.[398] AI is an art not a science, they say, and engineers do not follow a blueprint[399] when they build systems. However, many professions, such as medicine,[400] accounting,[401] and law,[402] think of their fields as art as much as science. "Professional knowledge, like all knowledge, is to some extent tacit; and it is this that gives the established professions their aura of mystery."[403] The field of

392. *Frequently Asked Questions*, AM. PSYCHIATRIC ASS'N, https://www.psychiatry.org/psychiatrists/practice/dsm/frequently-asked-questions [https://perma.cc/C9MA-XLN7].

393. AM. PSYCHIATRIC ASS'N, DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS at 6–7 (5th ed. 2013).

394. *Submit Proposals for Making Changes To DSM-5-TR*, AM. PSYCHIATRIC ASS'N, https://www.psychiatry.org/psychiatrists/practice/dsm/submit-proposals [https://perma.cc/A5DT-KYPY].

395. *View and Comment on Recently Proposed Changes to DSM-5-TR*, AM. PSYCHIATRIC ASS'N, https://www.psychiatry.org/psychiatrists/practice/dsm/proposed-changes [https://perma.cc/F7CG-DV9E].

396. *Plan For Federal AI Standards Engagement,* NIST (Apr. 5, 2022), https://www.nist.gov/artificial-intelligence/plan-federal-ai-standards-engagement [https://perma.cc/772P-UYLT].

397. *ACM Code of Ethics and Professional Conduct*, ACM (2018), https://www.acm.org/code-of-ethics [https://perma.cc/FTP7-4ZB3].

398. Hadrien Pouget, *The EU's AI Act Is Barreling Toward AI Standards that Do Not Exist*, LAWFARE (Jan. 12, 2023, 8:16 AM), https://www.lawfaremedia.org/article/eus-ai-act-barreling-toward-ai-standards-do-not-exist [https://perma.cc/MY76-YPNR].

399. Noam Hassenfeld, *Even the Scientists Who Build AI Can't Tell You How It Works*, VOX (July 15, 2023, 6:00 AM), https://www.vox.com/unexplainable/2023/7/15/23793840/chat-gpt-ai-science-mystery-unexplainable-podcast [https://perma.cc/KC4D-SG7X].

400. Robert Pearl, *Medicine Is an Art, Not a Science: Medical Myth or Reality?*, FORBES (June 12, 2014, 12:55 PM), https://www.forbes.com/sites/robertpearl/2014/06/12/medicine-is-an-art-not-a-science-medical-myth-or-reality/ [https://perma.cc/6A36-XL3J].

401. Robert R. Sterling, *Toward a Science of Accounting*, 31 FIN. ANALYSTS J. 28 (1975).

402. Basha Rubin, *Is Law an Art or a Science?: A Bit of Both*, FORBES (Jan. 13, 2015, 1:07 PM), https://www.forbes.com/sites/basharubin/2015/01/13/is-law-an-art-or-a-science-a-bit-of-both/ [https://perma.cc/QV29-S2R9].

403. Wilensky, *supra* note 291, at 149 (explaining that "'there are things that we know but cannot tell': the doctor's recognition of the characteristic appearance of a disease, the taxonomist's recognition of the specimen of a species").

medicine does not always know why pharmaceutical drugs work[404]—but they still use them. The field of accounting has had to adapt on the fly to cryptocurrency[405] as law had to adapt to generative AI.[406] There can be infinite permutations of circumstances impossible to address in ex ante practice guidelines. No profession has achieved perfect and complete standards that reduce the likelihood of error to zero—but that does not undermine the value of setting a floor, prohibiting the worst behavior, and encouraging the adoption of whatever techniques the field has developed to minimize harm, even if just a bit.

Finally, to maintain professional integrity, AI engineers must discipline irresponsible members of their own community. When AI engineers are rewarded, not punished, by employers for irresponsible practices, and courts are resistant to recognize AI harms, disciplinary tribunals are necessary to cast off noncompliant members of the community. In medicine, this work is done by state licensing boards[407] and some disciplinary actions are creatures of state law.[408] The board reviews and investigates complaints, escalating some to formal hearings and providing the opportunity for all parties to appeal.[409] Penalties can include revocation of a license, suspension of a license, license limitations to a specific area or type of work, education or training requirements, censure, fines, community service, and probationary monitoring.[410]

The professionalization process could take years, but that is true of most regulatory interventions. Professionalization pulls from a familiar playbook, spurring the field to elevate its standard of care without reinventing the wheel.

404. Carolyn Y. Johnson, *One Big Myth About Medicine: We Know How Drugs Work*, WASH. POST (July 23, 2015, 2:00 PM), https://www.washingtonpost.com/news/wonk/wp/2015/07/23/one-big-myth-about-medicine-we-know-how-drugs-work/ [https://perma.cc/CQG5-VTJ2].

405. Andrew Lom, Todd Schroeder, Susan Linda Ross, Rachael Hashmall & Hersh Verma, *IRS Releases First Cryptocurrency Guidance in Five Years*, NORTON ROSE FULBRIGHT (Nov. 2019), https://www.nortonrosefulbright.com/en-us/knowledge/publications/e29130a9/irs-releases-first-cryptocurrency-guidance-in-five-years [https://perma.cc/G946-N5RS].

406. Sam Tobin, *English Judges Get First-Ever Guidance on Artificial Intelligence*, REUTERS (Dec. 11, 2023, 8:38 PM), https://www.reuters.com/legal/transactional/english-judges-get-first-ever-guidance-artificial-intelligence-2023-12-12/ [https://perma.cc/4YK9-RZW5].

407. *See, e.g.*, *Physician and Physician Assistants Disciplinary and Other Actions*, N.Y. STATE DEP'T HEALTH (July 2012), https://www.health.ny.gov/professionals/doctors/conduct/index.htm [https://perma.cc/2S32-448M].

408. *See, e.g.*, *Relevant New York State Laws*, N.Y. STATE DEP'T HEALTH (Sept. 2019), https://www.health.ny.gov/professionals/doctors/conduct/laws.htm [https://perma.cc/ETF6-QXPB].

409. *Understanding New York's Medical Conduct Program - Physician Discipline*, *supra* note 386.

410. *Id.*

B. Countering the Risk of Professional Protectionism

Professionalization is not without its discontents and not all of their complaints are specious. Opponents to professionalization argue it does not improve quality of care while increasing barriers to entry, reducing overall public welfare.[411]

Yes, the justifications for licensing natural hair braiding, massage therapist, barber, and cemetery associate broker are dubious at best.[412] And some licensing requirements can seem nonsensical—"[c]osmetologists, for example, are required, on average, to have ten times as many days of training as Emergency Medical Technicians (EMT) must have."[413] These realities and the fact that today, nearly a third of the American workforce works in a licensed profession[414] suggests the system is being weaponized to achieve ulterior motives.

While there are many licensing regimes that stink of pretext, that is not an indictment against the whole institution. For one, allegations that licensing regimes are anticompetitive, hurt consumers, and do not improve the quality of service are "not sufficiently theorized, justified, or empirically grounded" to support deregulatory interventions.[415] Anecdotal "horror stories" of outlier cases do not constitute proof.[416] Further, even licensing's harshest critics concede that not all licensing regimes are harmful and that "[s]ome no doubt improve service quality and public safety enough to justify the costs,"[417] especially when "consumers cannot evaluate the quality of a professional's services."[418] No one has foresworn licensing entirely.

That being said, some concerns critics raise are not invalid. While they do not justify abandoning this solution altogether, they do caution care. To ensure professionalization works well, the system must be designed to guard against professional protectionism. This protectionism manifests in three core ways: (1) lowering the standard of care; (2) underenforcing standards;

---

411. *See generally* MORRIS M. KLEINER, THE HAMILTON PROJECT, REFORMING OCCUPATIONAL LICENSING POLICIES (2015), https://www.hamiltonproject.org/wp-content/uploads/2023/01/reforming_occupational_licensing_morris_kleiner_final.pdf [https://perma.cc/MD3S-ERW8].

412. PA. DEP'T OF STATE, 50 STATE COMPARISON REPORT: A COMPARISON OF STATE OCCUPATIONAL LICENSURE REQUIREMENTS AND PROCESSES 13 (2021), https://www.pa.gov/content/dam/copapwp-pagov/en/dos/resources/professional-licensing/50-state-reports/50-State-Comparison-Report-full.pdf [https://perma.cc/D3KJ-S3YQ].

413. Edlin & Haw, *supra* note 42, at 1096–97.

414. Scheffler, *supra* note 385, at 306.

415. Sandeep Vaheesan & Frank Pasquale, *The Politics of Professionalism: Reappraising Occupational Licensure and Competition Policy*, 14 ANN. REV. L. & SOC. SCI. 309, 312 (2018) ("Commentators all too often extrapolate from horror stories to make claims about the entirety of licensing . . . .").

416. *Id.*

417. Edlin & Haw, *supra* note 42, at 1098.

418. COX & FOSTER, *supra* note 217, at vi; *see also id.* at v.

and (3) erecting unjustified barriers to entry. Avoiding these pitfalls is more than possible—it just requires heeding lessons from history.

Medical professionals lowered the bar against which they are measured by ceding control over standards to state legislatures. The "pendulum swung back from the heyday of professional power that began in the late 19th century and reached its apogee in the 1950s and 1960s" when the molehead of alternative medicine reemerged and physicians began to lose favor with the public and courts again, allowing insurance to hijack medical standards of care.[419] They sunk the ship rather than hand it to pirates, inviting state legislatures to take control over medical standards.[420] Legislatures ossified practice guidelines doomed to be outdated by encoding them in statutes that "plaintiffs could not challenge . . . as substandard."[421] To further guard against a resurgence of litigation, legislatures adopted tort reforms "sheltering the healthcare industry from scrutiny and accountability, further reducing incentives for safe care," such as damage caps, limits on plaintiff's recovery for attorney's fees,[422] and reduced insurance policy coverage requirements.[423]

These outcomes are preventable. For one, diligent enforcement of socially conscious standards would avoid the public disdain that swells into waves of litigation.[424] Two, the federal government can resist professional efforts to capture the legislative process, forbearing from adopting statutory standards of care. Three, the federal government can keep the insurance industry in check, shielding AI engineers from pressure to compromise standards for economic reasons.[425]

Due to resource constraints, licensing boards underenforce standards, evincing a tolerance for low quality medical care.[426] State medical boards are "understaffed, overworked and poorly funded"[427] and "simply do not

419. Mehlman, *supra* note 308, at 1187; *see also id.* at 1188.

420. Pearson, *supra* note 335, at 557.

421. Mehlman, *supra* note 308, at 1195; *see also id.* at 1193–96.

422. SAKS & LANDSMAN, *supra* note 239, at 2; *see also id.* at 15; BERNARD S. BLACK, DAVID A. HYMAN, MYUNGHO PAIK, WILLIAM M. SAGE & CHARLES SILVER, MEDICAL MALPRACTICE LITIGATION 33–47 (2021).

423. Charles Silver, David A. Hyman, Bernard S. Black & Myungho Paik, *Policy Limits, Payouts, and Blood Money: Medical Malpractice Settlements in the Shadow of Insurance*, 5 U.C. IRVINE L. REV. 559, 585 (2015) (explaining that lower policy limits led to less victim compensation).

424. *See* Pearson, *supra* note 335.

425. Mehlman, *supra* note 308, at 1187–88.

426. *See, e.g.*, Nadia N. Sawicki, *Character, Competence, and the Principles of Medical Discipline*, 13 J. HEALTH CARE L. & POL'Y 285, 290 (2010); Deborah L. Rhode, *The Profession and the Public Interest*, 54 STAN. L. REV. 1501, 1512 (2002).

427. Jurecic, *supra* note 368.

have the resources" to monitor physician practices.[428] This likely contributes to the fact that the public often sees professional disciplinary actions as "too slow, too secret, too soft, and too self-regulated."[429] As true for professional enforcement as it is for AI systems: opacity erodes confidence.[430]

The government can ensure the success of professional disciplinary actions by amply resourcing the federal licensing board and providing transparency into the disciplinary process. By strategically appointing lay licensing board members with a distinct interest in enforcing standards and granting standards enforcement authority to a separate agency, such as the Federal Trade Commission (FTC), the government can overcome concerns that AI engineers won't discipline their own.[431] While the medical profession is largely self-regulated, accounting standards are enforced by the Securities and Exchange Commission.[432]

Physician control over licensing boards have allowed them to erect rules limiting entry to the profession.[433] By requiring that a majority of the board consist of practicing professionals and obviating the influence of lay board members, statutes governing licensing bodies enable anticompetitive behavior.[434] This has led to licensure laws that "hinder access to health care" by preserving scarcity of physicians[435] while mandating their involvement in aspects of patient care for which they are not required.[436]

The government already has, and can continue to, actively combat the professional capture of licensing boards.[437] The Supreme Court has already

428. Rob Kuznia, Scott Bronstein, Curt Devine & Drew Griffin, *They Take an Oath to Do No Harm, but These Doctors Are Spreading Misinformation About the Covid Vaccine*, CNN (Oct. 20, 2021, 4:07 PM), https://www.cnn.com/2021/10/19/us/doctors-covid-vaccine-misinformation-invs/index.html [https://perma.cc/R3ST-ZP2E].

429. Rhode, *supra* note 426, at 1512.

430. Bruce A. Green, *Selectively Disciplining Advocates*, 54 CONN. L. REV. 151, 193 (2022).

431. *See* Kathleen Leslie et al., *Regulating Health Professional Scopes of Practice: Comparing Institutional Arrangements and Approaches in the US, Canada, Australia and the UK*, 19 HUM. RES. FOR HEALTH, Jan. 28, 2021, at 3 (discussing the need for external oversight to prevent professional regulatory capture, the role of centralized review boards in maintaining public accountability, and the inclusion of lay members to ensure transparency and prevent anti-competitive practices in self-regulation processes).

432. Andrew F. Tuch, *The Self-Regulation of Investment Bankers*, 83 GEO. WASH. L. REV. 101, 120 (2014).

433. Edlin & Haw, *supra* note 42, at 1095–96, 1108.

434. Allensworth, *supra* note 42, at 1570–71.

435. Scheffler, *supra* note 385, at 312.

436. Edlin & Haw, *supra* note 42, at 1097.

437. Scheffler, *supra* note 385, at 321; *see also id.* at 298–99 ("For instance, the federal government has recently eased licensing restrictions for health care providers in certain areas where it already possesses regulatory authority, created incentives for states and professional bodies to experiment with reforms, intensified its focus on licensing boards' anti-competitive conduct, and created additional pressure for state-level reforms through expanding health insurance and promoting delivery system reforms under the Affordable Care Act (ACA).").

set the groundwork[438] for vigorous FTC enforcement of antitrust laws against licensing boards.[439] The apparatus to check any co-opting of licensing requirements for anticompetitive purposes already exists. The government can also take a more active role in supervising licensing boards[440] and it can design the licensing board to have as many non-practicing AI engineers as practicing ones.

CONCLUSION

Like the internet itself, AI is here to stay. The task at hand sounds deceptively simple: maximize its benefits, minimize its harm. Figuring out how to accomplish this, however, boggles the mind. It requires the capacity to distinguish good AI from bad AI; to find a needle in a black box. Failure to do so can either deny society lifesaving solutions to problems that have plagued us for centuries or doom society to AI-driven cataclysm. Unfortunately, companies lack the incentive to build safe, secure, and trustworthy AI; they are in the midst of an AI arms race. And our usual levers of influence—substantive regulation and legal action—are insufficient, on their own, to get the job done.

The fact of the matter is: none among us have much experience reckoning with this new technology—those who do are the ones building it. The best way to harness the experience and expertise of AI engineers in navigating our new world order is to enlist their help in creating it. Accordingly, this Article argues that the best way to reconcile humanity's wellbeing with the awesome power of AI is to conscript AI engineers to the task by professionalizing them. They know the technology best and are best positioned to anticipate how it can cause harm—and fix it. Professionalization would divorce the interests of individual engineers from the corporate incentives driving our current AI race to the bottom and imbue them with social responsibility. Forced to consider the public interest on pain of losing their livelihood, qualified and licensed AI engineers are the best vanguard society can hope for in the age of AI.

While the process would take years, and presents risks of its own kind, that is true of any regulatory intervention. Best start with the approach that might actually work.

---

438. N.C. State Bd. of Dental Exam'rs v. FTC, 574 U.S. 494, 503–04 (2015) (holding that state licensing boards could be subject to antitrust claims and that immunity only applied if the alleged anticompetitive practice was "clearly articulated and affirmatively expressed as state policy," and that "the policy . . . [be] actively supervised by the State" (quoting FTC v. Phoebe Putney Health Sys., Inc., 568 U.S. 216, 225 (2013))).

439. Scheffler, *supra* note 385, at 331.

440. *See* Allensworth, *supra* note 42, at 1602.