LAW-FOLLOWING AI: DESIGNING AI AGENTS TO OBEY HUMAN LAWS

Cullen O'Keefe, Ketan Ramakrishnan,** Janna Tay*** & Christoph Winter*****

Artificial intelligence (AI) companies are working to develop a new type of actor: "AI agents," which we define as AI systems that can perform computer-based tasks as competently as human experts. Expert-level AI agents will likely create enormous economic value but also pose significant risks. Humans use computers to commit crimes, torts, and other violations of the law. As AI agents progress, therefore, they will be increasingly capable of performing actions that would be illegal if performed by humans. Such lawless AI agents could pose a severe risk to human life, liberty, and the rule of law.

Designing public policy for AI agents is one of society's most important tasks. With this goal in mind, we argue for a simple claim: in high-stakes deployment settings, such as government, AI agents should be designed to rigorously comply with a broad set of legal requirements, such as core parts of constitutional and criminal law. In other words, AI agents should be loyal to their principals, but only within the bounds of the law: they should be designed to refuse to take illegal actions in the service of their principals. We call such AI agents "Law-Following AIs" (LFAI).

The idea of encoding legal constraints into computer systems has a respectable provenance in legal scholarship. But much of the existing scholarship relies on outdated assumptions about the (in)ability of AI systems to reason about and comply with open-textured, natural-language laws. Thus, legal scholars have tended to imagine a process of "hard-coding" a

* Director of Research, Institute for Law & AI; Research Affiliate, Centre for the Governance of AI.

** Associate Professor of Law, Yale Law School.

*** Research Scholar, Institute for Law & AI.

**** Assistant Professor of Law and AI, University of Cambridge; Director, Institute for Law & AI; Research Affiliate, Harvard University. For useful comments and discussions, the authors thank Yonathan Arbel, Jack Boeglin, Nick Caputo, Stephen Casper, Alan Chan, Rebecca Crootof, Kevin Frazier, Gillian Hadfield, Dan Hendrycks, Noam Kolt, Matthijs Maas, Richard Ngo, Alan Rozenshtein, Peter Salib, Chinmayi Sharma, Carl Shulman, Helen Toner, Aaron Tucker, Peter Wills, and discussants at the Center for Security and Emerging Technology at Georgetown University. For help editing this Article, we thank Suzanne Van Arsdale and the editors of the *Fordham Law Review*.

small number of specific legal constraints into AI systems by translating legal texts into formal machine-readable computer code. Existing frontier AI systems, however, are already competent at reading, understanding, and reasoning about natural-language texts, including laws. This development opens new possibilities for their governance.

Based on these technical developments, we propose aligning AI systems to a broad suite of existing laws as part of their assimilation into the human legal order. This would require directly imposing legal duties on AI agents. While this would be a significant change to legal ontology, it is both consonant with past evolutions (such as the invention of corporate personhood) and consistent with the emerging safety practices of several leading AI companies.

This Article aims to catalyze a field of technical, legal, and policy research to develop the idea of law-following AI more fully. It also aims to flesh out LFAI's implementation so that our society can ensure that widespread adoption of AI agents does not pose an undue risk to human life, liberty, and the rule of law. Our account and defense of law-following AI is only a first step and leaves many important questions unanswered. But if the advent of AI agents is anywhere near as important as the AI industry supposes, then law-following AI may be one of the most neglected and urgent topics in law today, especially in light of increasing governmental adoption of AI.

"[A] code of cyberspace, defining the freedoms and controls of cyberspace, will be built. About that there can be no debate. But by whom, and with what values? That is the only choice we have left to make."[1]

"AI is highly likely to be the control layer for everything in the world. How it is allowed to operate is going to matter perhaps more than anything else has ever mattered."[2]

INTRODUCTION

The law, as it exists today, aims to benefit human societies by structuring, coordinating, and constraining human conduct. Even where the law recognizes artificial legal persons—such as sovereign entities and corporations—it regulates them by regulating the human agents through

1. LAWRENCE LESSIG, CODE: VERSION 2.0, at 6 (2006).
2. Marc Andreessen, *Why AI Will Save the World*, MARC ANDREESSEN SUBSTACK (June 6, 2023), https://pmarca.substack.com/p/why-ai-will-save-the-world [https://perma.cc/463R-G3DZ].

which they act.[3] Proceedings in rem really concern the legal relations between humans and the res.[4] Animals may act, but their actions cannot violate the law;[5] the premodern practice of prosecuting them thus mystifies the modern mind.[6] To be sure, the law may protect the *interests* of animals and other nonhuman entities, but it invariably does so by imposing *duties* on humans.[7] Our modern legal system, at bottom, always aims its commands at human beings.

But technological development has a pesky tendency to challenge long-held assumptions upon which the law is built.[8] Frontier AI developers such as OpenAI, Anthropic, Google DeepMind, and xAI are starting to release the first agentic AI systems: AI systems that can do many of the things that humans can do through a computer, such as navigating the internet, interacting with counterparties online, and writing software.[9] Today's agentic AI systems are still brittle and unreliable in various respects.[10] These technical limitations also limit the impact of today's AI agents. Accordingly, today's AI agents are not our primary object of concern. Rather, our proposal targets the *fully capable* AI agents that AI companies seek to build: AI systems "that can do anything a human can do in front of a computer"[11] as competently as a human expert. Given the generally rapid rate of progress in advanced AI over the past few years,[12] the

3. *See infra* Part III.B.2; *see also* Richard L. Cupp, Jr., *Litigating Nonhuman Animal Legal Personhood*, 50 TEX. TECH L. REV. 573, 591–92 (2018).

4. *See, e.g.*, Morris E. Cohn, *Jurisdiction in Actions in Rem and in Personam*, 14 ST. LOUIS L. REV. 170, 171 (1929) ("An action *in rem* is one whose judgment is an official decree of the status of a thing *as it concerns persons*. It is binding on every interested party." (emphasis added)).

5. *E.g.*, People *ex rel.* Nonhuman Rts. Project, Inc. v. Lavery, 998 N.Y.S.2d 248, 251 (App. Div. 2014) ("Needless to say, unlike human beings, chimpanzees cannot bear any legal duties, submit to societal responsibilities or be held legally accountable for their actions.").

6. *See, e.g.*, Peter Dinzelbacher, *Animal Trials: A Multidisciplinary Approach*, 32 J. INTERDISC. HIST. 405, 405 (2002).

7. *See infra* Part II.A (discussing "quasi-persons").

8. *See* Rebecca Crootof & BJ Ard, *Structuring Techlaw*, 34 HARV. J.L. & TECH. 347, 414 (2021) ("As with individual laws, entire legal regimes can be grounded on assumptions rendered inaccurate by technological development.").

9. *See generally infra* Part I.A.

10. For an instructive (albeit somewhat outdated) discussion of the shortcomings of current AI agents, see Sayash Kapoor, Benedikt Stroebl, Zachary S. Siegel, Nitya Nadgir & Arvind Narayanan, AI Agents That Matter (July 2, 2024) (unpublished manuscript), https://arxiv.org/pdf/2407.01502 [https://perma.cc/7J8E-E6T5].

11. *ACT-1: Transformer for Actions*, ADEPT (Sep. 14, 2022), https://www.adept.ai/blog/act-1 [https://perma.cc/K878-WZTX]. The concept of an AI agent is explained more fully *infra* Part I.A. While we use this definition for the purposes of motivating LFAI, we think that a less demanding definition is likely desirable for legal and policy purposes. *See* Question 1 of the Research Agenda, *infra* Part VI. Accordingly, where the distinction is important, we sometimes call AI agents that meet or exceed human expert performance on literally all computer-based tasks "full AI agents" or "fully capable AI agents."

12. *See generally AI Benchmarking Hub*, EPOCH AI, https://epoch.ai/data/ai-benchmarking-dashboard [https://perma.cc/K3QV-QR27] (last visited Mar. 13, 2025).

biggest AI companies might achieve this goal much sooner than many outside of the AI industry expect.[13]

If AI companies succeed at building fully capable AI agents (hereinafter simply "AI agents")—or come anywhere close to succeeding—the implications will be profound. A dramatic expansion in supply of competent virtual workers could supercharge economic growth and dramatically improve the speed, efficiency, and reliability of public services.[14] But AI agents could also pose a variety of risks, such as precipitating severe economic inequality and dislocation by reducing the demand for human cognitive labor.[15] These economic risks deserve serious attention.

Our focus in this Article, however, is on a different set of risks: risks to life, liberty, and the rule of law. Many computer-based actions are crimes, torts, or otherwise illegal. Thus, sufficiently sophisticated AI agents could engage in a wide range of behavior that would be illegal if done by a human, with consequences that are no less injurious.[16]

These risks might be particularly profound for AI agents cloaked with state power. If they are not designed to be law following,[17] government AI agents may be much more willing to follow unlawful orders, or use unlawful methods to accomplish their principals' policy objectives, than human government employees.[18] A government staffed largely by non-law-following AI agents (what we call "AI henchmen")[19] would be a government much more prone to abuse and tyranny.[20] As the federal

13. On the general phenomenon of AI capabilities often progressing rapidly and discontinuously, see generally Markus Anderljung et al., Frontier AI Regulation: Managing Emerging Risks to Public Safety 10–13 (Nov. 7, 2023) (unpublished manuscript), https://arxiv.org/pdf/2307.03718 [https://perma.cc/JF28-M5QT]. Frontier AI systems are already very good at coding and improving rapidly. *Compare* Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press & Karthik Narasimhan, *SWE-Bench: Can Language Models Resolve Real-World Github Issues?*, PROC. 12TH INT'L CONF. ON LEARNING REPRESENTATIONS (2024), https://openreview.net/pdf?id=VTF8yNQM66 [https://perma.cc/8SXZ-KEF3] (best model could solve 1.9 percent of real-world coding problems), *with* SWE-BENCH, https://www.swebench.com/#verified [https://perma.cc/MH67-G58B] (last visited Mar. 18, 2025) (current leading AI system on human-validated version of the same benchmark solves 64.6 percent of problems). Since the design of AI agents is largely a coding task, AI systems that can code very well could drive rapid improvements in AI agent capabilities. *See generally, e.g.*, DANIEL ETH & TOM DAVIDSON, WILL AI R&D AUTOMATION CAUSE A SOFTWARE INTELLIGENCE EXPLOSION? (2025), https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion.pdf [https://perma.cc/NQ9W-EG9D].

14. *See infra* Part I.B.

15. *See infra* Part I.B (noting that AI agents could perform most core job tasks of many occupations).

16. *See infra* Part I.D.

17. We will show below that the AI alignment literature implies that AI agents may not be law-following by default. *See infra* Part IV.A.

18. *See infra* Part III.C.1.

19. *See infra* Part I.C.

20. *See generally, e.g.*, TOM DAVISON, LUKAS FINNVEDEN & ROSE HADSHAR, AI-ENABLED COUPS: HOW A SMALL GROUP COULD USE AI TO SEIZE POWER (2025), https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power.pdf

government lays the groundwork for the eventual automation of large swaths of the federal bureaucracy,[21] those who care about preserving the American tradition of ordered liberty must develop policy frameworks that anticipate and mitigate the new risks that such changes will bring.

This Article is our contribution to that project. We argue that the law should impose a broad array of legal duties on AI agents—of similar breadth to the legal obligations applicable to humans—to blunt the risks from lawless AI agents. We argue, moreover, that the law should require AI agents to be *designed*[22] to rigorously satisfy those duties.[23] We call such agents *Law-Following AIs* (LFAI).[24] We also use "LFAI" to denote our policy proposal: ensuring that AI agents are law following.

To some, the idea that AI should be designed to follow the law may sound absurd. To others, it may sound obvious.[25] Indeed, the idea of designing AI systems to obey some set of laws has a long provenance, going back to Isaac Asimov's (in)famous[26] Three Laws of Robotics.[27] But our vision for LFAI differs substantially from much of the existing legal scholarship on the automation of legal compliance. Much of this existing scholarship envisions the design of law-following computer systems as a process of hard-coding a small, fixed, and formally-specified set of decision rules into the code of a computer system prior to its deployment in order to address foreseeable

---

[https://perma.cc/J6DL-EBE6]; Justin B. Bullock, Samuel Hammond & Seb Krier, AGI, Governments, and Free Societies (Mar. 13, 2025) (unpublished manuscript), https://arxiv.org/pdf/2503.05710 [https://perma.cc/Q5MY-A9FX] (arguing that the creation of advanced AI risks "pushing societies toward . . . a 'despotic Leviathan' through enhanced state surveillance and control").

21. *See, e.g.*, Jeff Stein, Elizabeth Dwoskin, Hannah Natanson & Jonathan O'Connell, *In Chaotic Washington Blitz, Elon Musk's Ultimate Goal Becomes Clear*, WASH. POST (Feb 8, 2025), https://www.washingtonpost.com/business/2025/02/08/doge-musk-goals/ (on file with the *Fordham Law Review*) ("'The end goal is replacing the human workforce with machines,' said a U.S. official closely watching DOGE activity. 'Everything that can be machine-automated will be. And the technocrats will replace the bureaucrats.'"). Other governments are contemplating similar efforts. *See, e.g.*, Rowena Mason, *AI Should Replace Some Work of Civil Servants, Starmer to Announce*, GUARDIAN (Mar. 12, 2025, 18:30 ET), https://www.theguardian.com/technology/2025/mar/12/ai-should-replace-some-work-of-civil-servants-under-new-rules-keir-starmer-to-announce [https://perma.cc/ST4E-4VVG].

22. The importance of design is discussed *infra* Part III.

23. The question of when, exactly, LFAI should be required is part of our broader research agenda. *See* Question 5 of the Research Agenda, *infra* Part VI.

24. In part, this Article develops and modifies ideas first informally introduced in posts compiled at Cullen O'Keefe, *Law-Following AI*, ALIGNMENT FORUM (Aug. 5, 2022), https://www.alignmentforum.org/s/ZytYxd523oTnBNnRT [https://perma.cc/QJ8M-J5JD].

25. *Cf.* Dylan Hadfield-Menell, McKane Andrus & Gillian K. Hadfield, *Legible Normativity for AI Alignment: The Value of Silly Rules*, *in* AIES'19: PROC. 2019 AAAI/ACM CONF. ON A.I., ETHICS, & SOC'Y 115, 115 ("[I]t has become commonplace to assert that autonomous agents will have to be built to follow human norms and laws.").

26. For a discussion of why the Three Laws of Robotics are generally not considered a serious AI safety proposal, see Computerphile, *Why Asimov's Laws of Robotics Don't Work*, YOUTUBE (Nov. 6, 2015), https://www.youtube.com/watch?v=7PKx3kS7f4A.

27. The Three Laws were first articulated in Isaac Asimov, *Runaround*, *in* I, ROBOT 25, 37 (2004).

classes of legal dilemmas.[28] Such discussions often assume that computer systems are unable to interpret, reason about, and comply with open-textured natural-language laws.[29]

AI progress has undermined that assumption. Today's frontier AI systems can already reason about existing natural-language texts, including laws, with some reliability—no translation into computer code required.[30] They can also use search tools to ground their reasoning in external, web-accessible sources of knowledge,[31] such as the evolving corpus of statutes and case law. Thus, the capabilities of existing frontier AI systems strongly suggest that future AI agents will be capable of the core tasks needed to follow natural-language laws, including finding applicable laws, reasoning about them, tracking relevant changes to the law, and even consulting lawyers in hard cases. Indeed, frontier AI companies are already instructing their AI agents to follow the law,[32] suggesting they believe that the development of law-following AI agents is *already* a reasonable goal.

A separate strand of existing literature seeks to prevent harms from highly autonomous AI agents by holding the principals (that is, developers, deployers, or users) of AI agents liable for legal wrongs committed by the agent, through a form of respondeat superior liability.[33] This would, in some sense, incentivize those principals to cause their AI agents to follow the law, at least insofar as the agents' harmful behavior can be thought of as law

28. *Cf.* SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS AGENTS 166 (2011) (discussing a law-following autonomous vehicle obeying traffic laws); Mark A. Chinen, *The Co-Evolution of Autonomous Machines and Legal Responsibility*, 20 VA. J.L. & TECH. 338, 379–80 (2016) ("[P]rogramming machines to obey the law is possible only to a certain extent: law cannot always be reduced to a set of rules of decision."); Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1311, 1370–78 (2019) ("A court can order a robot, say, not to take race into account in processing an algorithm. . . . Someone will have to translate that injunction, written in legalese, into code the robot can understand."); Ronald Leenes & Federica Lucivero, *Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design*, 6 L. INNOVATION & TECH. 193, 218 (2014) ("[T]here is a clash between the inflexibility of code and flexibility and openness to interpretation of the law, which often requires creative reasoning."); Ashley Deeks, *Coding the Law of Armed Conflict: First Steps* 2 (Univ. of Va. Sch. of L., Pub. L. & Legal Theory Paper Series, Working Paper No. 2020-49), https://ssrn.com/abstract=3612329 (on file with the *Fordham Law Review*) ("It is surely true that lethal autonomous systems will be hard to engineer with embedded legal safeguards because their programmers necessarily must encode all applicable rules of [laws of armed conflict (LOAC)] into them; there will be no human to intercede between system decision and execution to ensure LOAC compliance."); SIMON CHESTERMAN, WE, THE ROBOTS?: REGULATING ARTIFICIAL INTELLIGENCE AND THE LIMITS OF THE LAW 230 (2021) ("A preliminary problem is that legal rules are typically expressed in natural language that may be difficult for a computer to parse.").

29. *See supra* note 28.

30. *See infra* Part I.E.1.

31. *See infra* Part I.E.1.

32. *See infra* Part I.E.2.

33. *See, e.g.*, Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043, 1065–67 (2020); CHOPRA & WHITE, *supra* note 28, at 24–25, 128–29.

breaking.[34] While we do not disagree with these suggestions, we think that our proposal can serve as a useful complement to them, especially in contexts where liability rules provide only a weak safeguard against serious harm. One important such context is government work, where immunity doctrines often protect government agents and the state from robust ex post accountability for lawless action.[35]

Combining these themes, we advocate that,[36] especially in such high-stakes contexts,[37] the law should require that AI agents be designed such that they have "a strong motivation to obey the law" as one of their "basic drives."[38] In other words, we propose not that specific legal commands should be hard-coded into AI agents (and perhaps occasionally updated),[39] but that AI agents should be designed to be law following in general.

To be clear, we do not necessarily advocate that AI agents must perfectly obey literally every law. Our claim is more modest in both scope and demandingness. While we are uncertain about which laws LFAIs should follow, adherence to some foundational laws—such as central parts of the criminal law, constitutional law, and basic tort law—seems much more important than adherence to more niche areas of law.[40] Moreover, we are open to the possibility that LFAIs should be permitted to run some amount of legal risk: that is, perhaps an LFAI should sometimes be able to take an action that, in its judgment,[41] may be illegal.[42] Relatedly, we think the case for LFAI is strongest in certain particularly high-stakes domains, such as when AI agents act as substitutes for human government officials or otherwise exercise government power.[43] We are unsure when LFAI requirements are justified in other domains.[44]

---

34. Respondeat superior liability applies only when the employee has committed a tort. *See* RESTATEMENT (THIRD) OF AGENCY § 2.04 (A.L.I. 2006). Accordingly, to apply respondeat superior to the principals of an AI agent, we need to be able to say that the behavior of the agent was tortious.

35. *See infra* Parts III.A, III.C.1, V.C.

36. For proposals closer to LFAI, see, e.g., Leon E. Wein, *The Responsibility of Intelligent Artifacts: Towards an Automation Jurisprudence*, 6 HARV. J.L. & TECH. 103 (1992); CHOPRA & WHITE, *supra* note 28, at 121, 145–50, 165–67; Elina Nerantzi & Giovanni Sartor, *"Hard AI Crime": The Deterrence Turn*, 44 OXFORD J. LEGAL STUD. 673 (2024); Ondrej Bajgar & Jan Horenovsky, *Negative Human Rights as a Basis for Long-Term AI Safety and Regulation*, 76 J.A.I. RSCH. 1043 (2023).

37. *See* Question 5 of the Research Agenda, *infra* Part VI (asking in which contexts LFAI should be mandated).

38. *See* CHOPRA & WHITE, *supra* note 28, at 165.

39. *See id.* at 166.

40. *See* Question 2 of the Research Agenda, *infra* Part VI (discussing which laws LFAIs should obey).

41. *See* Question 4 of the Research Agenda, *infra* Part VI (discussing how an LFAI should determine what its legal obligations are under legal uncertainty).

42. *See* Question 7 of the Research Agenda, *infra* Part VI (discussing how rigorously LFAIs should obey those laws by which they are bound).

43. *See infra* Parts III.C.1, V.C.

44. *See* Question 5 of the Research Agenda, *infra* Part VI (asking in which contexts LFAI should be mandated).

The remainder of this Article will motivate and explain the LFAI proposal in further detail. In Part I, we offer background on AI agents. We explain how AI agents could break the law and the risks to human life, liberty, and the rule of law this could entail. We contrast LFAIs with AI henchmen: AI agents that are loyal to their principals but take a purely instrumental approach to the law and are thus willing to break the law for their principal's benefit when they think they can get away with it. We note that, by default, there may be a market for AI henchmen. We also survey the legal reasoning capabilities of today's large language models and existing trends toward something like LFAI in the AI industry.

Part II provides the foundational legal framework for LFAI. We propose that the law treat AI agents as legal actors, which we define as entities on which the law imposes duties, even if they possess no rights of their own. Accordingly, we do *not* argue that AI agents should be legal persons. Our argument is narrower: because AI agents can comprehend laws, reason about them, and attempt to comply with them, the law should require them to do so. We also anticipate and address an objection that imposing duties on AI agents is objectionably anthropomorphic.

If the law imposes duties on AI agents, this leaves open the question of how to make AI agents comply with those duties. Part III answers this question as follows: AI agents should be *designed to* follow applicable laws, even when they are instructed or incentivized by their human principals to do otherwise. Our case for regulation through the design of AI agents draws on Professor Lawrence Lessig's insight that digital artifacts can be designed to achieve regulatory objectives.[45] Since AI agents are human-designed artifacts, we should be able to design them to refuse to violate certain laws in the first place.

Part IV observes that designing LFAIs is an example of AI alignment: the pursuit of AI systems that rigorously comply with constraints imposed by humans. We therefore connect insights from AI alignment to the concept of LFAI. We also argue that, in a democratic society, LFAI is an especially attractive and tractable form of AI alignment, given the legitimacy of democratically enacted laws.

Part V briefly explores how a legal duty to ensure that AI agents are law following might be implemented. We first note that ex post sanctions, such as tort liability and fines, can disincentivize the development, possession, deployment, and use of AI henchmen in many contexts. However, we also argue that ex ante regulation would be appropriate in some high-stakes contexts, especially government. Concretely, this would mean something like requiring a person who wishes to deploy an AI agent in a high-stakes context to demonstrate that the agent is law following prior to receiving permission to deploy it. We also consider other mechanisms that might help promote the adoption of LFAIs, such as nullification rules and technical

45. *See generally* LESSIG, *supra* note 1.

mechanisms that prevent AI henchmen from using large-scale computational infrastructure.

Our goal in this Article is to start, not end, a conversation about how AI agents can be integrated into the human legal order. Accordingly, we do not answer many of the important questions—conceptual, doctrinal, normative, and institutional—that our proposal raises. In Part VI, we articulate an initial research agenda for the design and implementation of a "minimally viable" version of LFAI. We hope that this research agenda will catalyze further technical, legal, and policy research to that end. If the advent of AI agents is anywhere near as significant as the AI industry claims, these questions may be among the most pressing in legal scholarship today.

I. AI AGENTS AND THE LAW

LFAI is a proposal about how the law should treat a particular class of future AI systems: AI agents.[46] In this part, we explain what AI agents are and how they could profoundly transform the world.

A. From Generative AI to AI Agents

The current AI boom began with advances in "generative AI": AI systems that create new content,[47] such as large language models (LLM). As the initialism suggests, these LLMs were originally limited to inputting and outputting text.[48] AI developers subsequently deployed "multimodal" versions of LLMs ("MLLM"),[49] such as OpenAI's GPT-4o[50] and Google's

46. Some readers might question why we focus on lawbreaking behavior by AI agents specifically. After all, other AI systems—including less agentic generative AI systems—can "violate the law" in some sense, such as by outputting illegal types of content. *See, e.g.*, Leslie Y. Garfield Tenzer, *Defamation in the Age of Artificial Intelligence*, 80 N.Y.U. ANN. SURV. AM. L. 135 (2024); Nina Brown, *Bots Behaving Badly: A Products Liability Approach to Chatbot-Generated Defamation*, 3 J. FREE SPEECH L. 389 (2023). It is true that generative AI systems can produce outputs that, if produced by humans, would usually be understood to violate the law. However, AI agents can by definition take a wider range of actions than generative AI systems and thereby threaten a wider range of legally protected interests. Ensuring that AI agents follow the law is therefore of broader interest and significantly higher stakes.

47. *E.g.*, *What Is Generative AI?*, MCKINSEY & CO. (Apr. 2, 2024), https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai [https://perma.cc/CBL8-RHLS].

48. *See, e.g.*, Alec Radford, Karthik Narasimhan, Tim Salimans & Ilya Sutskever, Improving Language Understanding by Generative Pre-Training (unpublished manuscript), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [https://perma.cc/M5BH-PRFQ] (last visited Feb. 4, 2025) (describing the model that would later be named GPT-1).

49. *See generally* Shukang Yin et al., *A Survey on Multimodal Large Language Models*, 11 NAT'L SCI. REV. 403 (2024). "Language" is still included in the "MLLM" initialism because the LLM foundation gives MLLMs reasoning capabilities that nonlanguage models lack. *See id.* at 1.

50. *See generally* OpenAI, GPT-4o System Card (Aug. 8, 2024) (unpublished manuscript), https://arxiv.org/pdf/2410.21276 [https://perma.cc/F32G-UXJX].

Gemini,[51] that can receive inputs and produce outputs in multiple modalities, such as text, images, audio, and video.

The core competency of generative AI systems is, of course, generating new content. Yet, the utility of generative AI systems is limited in crucial ways. Humans do far more on computers than generating text and images.[52] Many of these computer-based tasks are better understood as *actions*, not content generation. And even those tasks that are largely generative, such as writing a report on a complicated topic, require the completion of active subtasks, such as searching for relevant terms, identifying relevant literature, following citation trees, arranging interviews, soliciting and responding to comments, paying for software, and tracking down copies of papers. If a computer-based AI system could do these active tasks, it could generate enormous economic value by making computer-based labor—a key input into many production functions—much cheaper.[53]

Advances in generative AI kindled hopes[54] that, if MLLMs could use computer-based tools in addition to generating content, we could produce a new type of AI system:[55] a computer-based[56] AI system capable of performing any task[57] that a human can do using a computer and doing so

51. *See generally* Sundar Pichai & Demis Hassabis, *Introducing Gemini: Our Largest and Most Capable AI Model*, GOOGLE (Dec. 6, 2023), https://blog.google/technology/ai/google-gemini-ai/ [https://perma.cc/XMS5-LEKD].

52. *See infra* Part I.B.

53. *See* Philip Trammell & Anton Korinek, *Economic Growth Under Transformative AI* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31815, 2023), http://www.nber.org/papers/w31815 (on file with the *Fordham Law Review*); Matthew Barnett, *The Economic Consequences of Automating Remote Work*, EPOCH AI (Jan. 10, 2024), https://epoch.ai/gradient-updates/consequences-of-automating-remote-work [https://perma.cc/8HUW-XHM4].

54. *See, e.g.*, Zane Durante et al., Agent AI: Surveying the Horizons of Multimodal Interaction 5 (Jan. 25, 2024) (unpublished manuscript), https://arxiv.org/pdf/2401.03568 [https://perma.cc/2BCR-8CB6] ("[T]the AI community is on the cusp of a significant paradigm shift, transitioning from creating AI models for passive, structured tasks to models capable of assuming dynamic, agentic roles in diverse and complex environments."); Sam Altman, *Reflections*, SAM ALTMAN (Jan. 6, 2025, 08:37), https://blog.samaltman.com/reflections [https://perma.cc/DB62-MVV8] ("We believe that, in 2025, we may see the first AI agents 'join the workforce' and materially change the output of companies.").

55. Foundational work on agentic AI systems based on LLMs includes Timo Schick et al., *Toolformer: Language Models Can Teach Themselves to Use Tools*, *in* NIPS '23: PROC. 37TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 68539 (2023), https://dl.acm.org/doi/10.5555/3666122.3669119 [https://perma.cc/X4C5-D28R]; Iason Gabriel et al., The Ethics of Advanced AI Assistants (Apr. 28, 2024) (unpublished manuscript), https://arxiv.org/pdf/2404.16244 [https://perma.cc/PC34-LFW9] (using "AI assistants" instead of "AI agents").

56. We focus here on tasks that can be conducted through computers because AI systems that interact directly with the physical world are part of a distinct field: robotics. There is a widely shared perception in the AI field that progress in robotics has been much slower than progress in tasks that can be accomplished using only a computer. *See, e.g.*, Melissa Heikkilä, *Is Robotics About to Have Its Own ChatGPT Moment?*, MIT TECH. REV. (Apr. 11, 2024), https://www.technologyreview.com/2024/04/11/1090718/household-robots-ai-data-robotics [https://perma.cc/4HN9-QD5D]. Focus on computer-based AI agents is thus a comparatively conservative focus, since it assumes progress in a narrower range of capabilities.

57. AI researchers have long been able to develop AI systems that can use computers to accomplish tasks in narrow domains, such as playing computer games. *See* Alan Chan et al.,

with the same level of competence as a human expert. This is the concept of a fully capable "computer-using agent":[58] what we are calling simply an "AI agent." Give an AI agent any task that can be accomplished using computer-based tools, and an AI agent will, by definition, do it as well as an expert human worker tethered to their desk.[59]

AI agents, so defined, do not yet exist, but they may soon. Some of the first functional demonstrations of first-party agentic AI systems have come online in the past few months. In October 2024, Anthropic announced that it had trained its Claude line of MLLMs to perform some computer-use tasks, thus supplying one of the first public demonstrations of an agentic model from a frontier AI lab.[60] In January 2025, OpenAI released a preview of its Operator agent.[61] Operating system developers are working to integrate existing MLLMs into their operating systems,[62] suggesting a possible pathway toward the widespread commercial deployment of AI agents.

It remains to be seen whether (and, if so, on what timescale) these existing efforts will bear lucrative fruit. Today's AI agents are primarily a research and development project, not a market-proven product. Nevertheless, with so many companies investing so much toward full AI agents, it would be prudent to try to anticipate risks that could arise if they succeed.[63]

Harms from Increasingly Agentic Algorithmic Systems, *in* FAccT '23: PROC. 2023 ACM CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 651, 655, https://dl.acm.org/doi/pdf/10.1145/3593013.3594033 [https://perma.cc/N7QW-T7QQ]. An agent is more general-purpose if it can perform tasks in a wider variety of environments. Thus, an AI system that can play many video games is more agentic than an AI system that can play only one video game, and an AI agent that could do any computer-based task would be more general-purpose still.

58. *Computer-Using Agent*, OPENAI (Jan. 23, 2025), https://openai.com/index/computer-using-agent/ [https://perma.cc/9HSF-87GJ].

59. *See supra* note 11 and accompanying text (defining "AI agent" for the purpose of this Article). On the number of tasks that could be automated by an AI agent, see, e.g., Jonathan I. Dingel & Brent Neiman, *How Many Jobs Can Be Done at Home?*, J. PUB. ECON., July 2020, at 1, 1 ("We find that 37% of jobs in the United States can be performed entirely at home."); Tyna Eloundou, Sam Manning, Pamela Mishkin & Daniel Rock, *GPTs Are GPTs: Labor Market Impact Potential of LLMs*, 384 SCI. 1306, 1306 (2024) ("[W]e estimate that roughly 1.8% of jobs could have over half their tasks affected by LLMs with simple interfaces and general training. When accounting for current and likely future software developments that complement LLM capabilities, this share jumps to just over 46% of jobs.").

60. *See Developing a Computer Use Model*, ANTHROPIC (Oct. 23, 2024), https://www.anthropic.com/news/developing-computer-use [https://perma.cc/9UHL-FTZB].

61. *Introducing Operator*, OPENAI (Jan. 23, 2025), https://openai.com/index/introducing-operator/ [https://perma.cc/3YC3-T7GH].

62. *See, e.g.*, Microsoft, *Full Keynote: Introducing Copilot+ PCs*, YOUTUBE (May 20, 2024), https://www.youtube.com/watch?v=aZbHd4suAnQ; *OpenAI and Apple Announce Partnership to Integrate ChatGPT into Apple Experiences*, OPENAI (June 10, 2024), https://openai.com/index/openai-and-apple-announce-partnership/ [https://perma.cc/HTP6-N9VW].

63. *See generally* Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177 (2024) (arguing for an anticipatory approach to AI law and policy).

B. The World of AI Agents

Fully capable and widely available AI agents would profoundly change society.[64] We cannot possibly anticipate all the issues that they would raise, nor could a single paper adequately address all such issues.[65] Still, some illustration of what a world with AI agents might look like is useful for gaining intuition about the dynamics that might emerge. This picture will doubtless be wrong in many particulars, but hopefully it will illustrate the general profundity of the changes that AI agents would bring.

A very large number of valuable tasks can be done by humans "in front of a computer."[66] If organizations decide to capitalize on this abundance of computer-based cognitive labor, AI agents could rapidly be charged with performing a large share of tasks in the economy, including in important sectors. AI scientist agents would conduct literature reviews, formulate novel hypotheses, design experimental protocols, order lab supplies, file grant applications, scour datasets for suggestive trends, perform statistical analyses, publish findings in top journals, and conduct peer review.[67] AI lawyer agents would field client intake, spot legal issues facing their client, conduct research on governing law, analyze the viability of the client's claims, draft memoranda and briefs, draft and respond to interrogatories, and prepare motions. AI intelligence analyst agents would collect and review data from multiple sources, analyze it, and report its implications up the chain of command. AI inventor agents would create digital blueprints and models of new inventions, run simulations, and order prototypes—and so on across many other sectors. The result could be a significant increase in the rate of economic growth.[68]

In short, a world with AI agents would be a world in which a new type of actor[69] would be available to perform cognitive labor, potentially at low cost and massive scale. By default, anyone who needed computer-based tasks done could "employ" an AI agent to do it for them. Most people would use this new resource for the better.[70] But many would not.

---

64. *See* Gabriel et al., *supra* note 55, at i.

65. For the most comprehensive overview of the issues surrounding widely available AI agents, see *id.*

66. *See supra* note 11 and accompanying text (defining "AI agent" for the purpose of this Article).

67. *Cf. Introducing Deep Research*, OPENAI (Feb. 2, 2025), https://openai.com/index/introducing-deep-research/ [https://perma.cc/L6XN-FK5D] (example of a system that can perform some core research tasks).

68. *See* Trammell & Korinek, *supra* note 53; Barnett, *supra* note 53.

69. *See infra* Part II.A.

70. *Cf.* Jeremy Howard, *AI Safety and the Age of Dislightenment*, FAST.AI (July 10, 2023), https://www.fast.ai/posts/2023-11-07-dislightenment.html [https://perma.cc/DV5E-MEX6] ("[M]ost people are not Bad Guys. Most people will use [powerful AI] models to create, and to protect.").

C. Loyal AI Agents, Law-Following AIs, and AI Henchmen

We can understand AI agents within the principal-agent framework familiar to lawyers and economists.[71] For simplicity, we will assume that there is a single human principal giving instructions to their AI agent.[72] Following typical agency terminology, we can say that an AI agent is loyal if it consistently acts for the principal's benefit according to their instructions.[73]

Even if an AI agent is designed to be loyal, other design choices will remain. Specifically, the developer of an AI agent must decide how the agent will act when it is instructed or incentivized to break the law in the service of its principal. This Article compares two basic ways loyal AI agents could respond in such situations. The first is the approach advocated by this Article: loyal AI agents that follow the law, or LFAIs.

The case for LFAI will be made more fully throughout this Article. But it is important to note that loyal AI agents are not guaranteed to be law following by default.[74] This is one of the key implications of the AI alignment literature, discussed in more detail in Part IV.A below. Thus, LFAIs can be contrasted with a second possible type of loyal AI agent: AI henchmen. AI henchmen take a purely instrumental approach to legal prohibitions: they act loyally for their principal and will break laws when doing so if such lawbreaking serves the principal's goals and interests.

A loyal AI henchman would not be a haphazard lawbreaker. Good henchmen have some incentive to avoid doing anything that could cause their principal to incur unwanted liability or loss. This gives them reason to avoid many violations of law. For example, if human principals were held liable for the torts of their AI agents under an adapted version of respondeat superior liability,[75] then an AI henchman would have some reason to avoid committing torts, especially those that are easily detectable and attributable.

---

71. *See, e.g.*, Chan et al., *supra* note 57, at 653–54; Dylan Hadfield-Menell & Gillian Hadfield, *Incomplete Contracting and AI Alignment*, *in* AIES'19: PROC. 2019 AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 417; Anthony Aguirre, Gaia Dempsey, Harry Surden & Peter B. Reiner, *AI Loyalty: A New Paradigm for Aligning Stakeholder Interests*, 1 IEEE TRANSACTIONS ON TECH. & SOC'Y 128 (2020); Sebastian Benthall & David Shekman, *Designing Fiduciary Artificial Intelligence*, *in* EAAMO '23: PROC. 3RD ACM CONF. ON EQUITY & ACCESS IN ALGORITHMS, MECHANISMS, & OPTIMIZATION (2023), https://doi.org/10.1145/3617694.3623230 (on file with the *Fordham Law Review*). Note that, drawing on the methodology of Professor Noam Kolt in *Governing AI Agents*, this Article generally "use[s] structures, principles, and vocabulary developed in the common law of agency in order to shed light on the challenges involved in governing AI agents" without necessarily advocating for the wholesale application of agency law to AI agents. Noam Kolt, *Governing AI Agents*, 101 NOTRE DAME L. REV. (forthcoming) (manuscript at 9), https://ssrn.com/abstract=4772956 (on file with the *Fordham Law Review*).

72. More complicated arrangements (e.g., multiple principals, subagents) are imaginable and perhaps more likely in reality.

73. *See* RESTATEMENT (THIRD) OF AGENCY § 8.01 (A.L.I. 2020).

74. *See infra* Part IV.A.

75. For sources proposing this, see *supra* note 33.

Even if respondeat superior did not apply, the principal's exposure to ordinary negligence liability, other sources of liability, or simple reputational risk might give the AI henchman reason to obey the law. Similarly, a good henchman will decline to commit many crimes simply because the risk-reward tradeoff is not worth it. This is the classic case of the drug smuggler who studiously obeys traffic laws: the risk to the criminal enterprise from speeding and getting caught with drugs obviously outweighs the benefit of quicker transportation times.

But these are only instrumental disincentives to break the law. Henchmen are not inherently averse to lawbreaking or robustly predisposed to refrain from it. If violating the law is in the principal's interest, all things considered, then an AI henchman will simply go ahead and violate the law. Since, in humans, compliance with law is induced both by instrumental disincentives and an inherent respect for the law,[76] AI agents that lack the latter may well be more willing to break the law than humans.

Criminal enterprises will be attracted to loyal AI agents for the same reasons that legitimate enterprises will: efficiency, scalability, multitask competence, and cost savings over human labor. But AI henchmen, if available, might be particularly effective lawbreakers as compared to human substitutes. For example, because AI henchmen do not have selfish incentives, they would be less likely to betray their principals to law enforcement (for example, in exchange for a plea bargain).[77] AI henchmen could have erasable memory,[78] which would reduce the amount of evidence available to law enforcement. They would lack the impulsivity, common in criminal offenders,[79] that often presents a serious operational risk to the larger criminal enterprise. They could operate remotely, across jurisdictional lines, behind layers of identity-obscuring software, and be meticulous about covering their tracks. Indeed, they might hide their lawbreaking activities even from their principal, thus allowing the principal to maintain plausible deniability and therefore insulate the principal from accountability.[80] AI henchmen may also be willing to bribe or intimidate legislators, law

76. *See generally* TOM TYLER, WHY PEOPLE OBEY THE LAW (2006).

77. *Cf.* CHOPRA & WHITE, *supra* note 28, at 21 ("Well-designed agents are arguably less likely to breach their fiduciary duties than human ones.").

78. OpenAI's ChatGPT has a "temporary chat" feature where messages are deleted from OpenAI's records after thirty days. *See Temporary Chat FAQ*, OPENAI, https://help.openai.com/en/articles/8914046-temporary-chat-faq [https://perma.cc/D3FL-75P3] (last visited Aug. 1, 2025).

79. *See, e.g.*, Travis C. Pratt & Francis T. Cullen, *The Empirical Status of Gottfredson and Hirschi's General Theory of Crime: A Meta-Analysis*, 38 CRIMINOLOGY 931 (2000); Alexander T. Vazsonyi, Jakub Mikuska & Erin L. Kelley, *It's Time: A Meta-Analysis on the Self-Control-Deviance Link*, 48 J. CRIM. JUST. 48 (2017).

80. As an analogy, middle management in corporations sometimes engages in lawbreaking behavior to meet the performance expectations of senior management, while simultaneously hiding that lawbreaking behavior from senior management to insulate them from liability. *See* J.S. Nelson, *Disclosure-Driven Crime*, 52 U.C. DAVIS L. REV. 1487, 1536–49 (2019).

enforcement officials, judges, and jurors.[81] They would be willing to fabricate or destroy evidence, possibly more undetectably than a human could.[82] They could use complicated financial arrangements to launder money and protect their principal's assets from creditors.[83]

Certainly, most people would prefer not to employ AI henchmen and would probably be horrified to learn that their AI agent seriously harmed others to benefit them. But those with fewer scruples would find the prospect of employing AI henchmen attractive:[84] many ordinary people might not mind if their agents cut a few legal corners to benefit them.[85] If AI henchmen were available on the market, then, we might expect a healthy demand for them. After all, from the principal's perspective, every inherent law-following constraint is a tax on the principal's goals. And if LFAIs provide less utility to consumers, developers will have less reason to create them. So, insofar as AI henchmen are available on the market, and in the absence of significant legal mechanisms to prevent or disincentivize their adoption, some people will choose henchmen over LFAIs. The next part explores the harms that might result from the availability of AI henchmen.

D. Mischief from AI Henchmen: Two Vignettes

Under our definition, AI agents "can do anything a human can do in front of a computer."[86] One of the things humans do in front of a computer is violate the law.[87] One obvious example is cybercrimes—"illegal activity

81. *See* Cullen O'Keefe, *Law-Following AI 2: Intent Alignment + Superintelligence → Lawless AI (By Default)*, ALIGNMENT F. (Apr. 28, 2022), https://www.alignmentforum.org/s/ZytYxd523oTnBNnRT/p/9aSi7koXHCakb82Fz [https://perma.cc/MLS3-RXEM].

82. *See id.*

83. *See id.*

84. Our concept of an AI henchman is similar to the concept of a "delegate": an actor "to whom people can outsource the execution of unethical behaviour." Nils Köbis, Jean-Francois Bonnefon & Iyad Rahwan, *Bad Machines Corrupt Good Morals*, 5 NATURE HUM. BEHAV. 679, 681 (2021). As Köbis et al. note:

> [M]ore often than not, people do not explicitly instruct the delegates to break ethical rules but instead merely define their desired outcome and turn a blind eye to the modalities of achieving this goal. Thereby, the remitter avoids direct contact with the victims and can willfully ignore any possible ethical rule violations. Moreover, if inflicted harm becomes apparent, blame and responsibility can be deflected to the delegate, which alleviates the guilt experienced.

Id. (endnotes omitted).

85. The tensions between constraints imposed by design and user preferences for freedom are noted in Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 VAND. J. ENT. & TECH. L. 133, 141–42 (2020).

86. *See supra* note 11 and accompanying text (defining "AI agent" for the purpose of this Article).

87. For now, we will use "violate the law" to mean "to take actions—or refrain from acting—in contravention of duties imposed by law." *Cf. Violation*, BLACK'S LAW DICTIONARY (12th ed. 2024) (defining "violation" as "the contravention of a right or duty").

carried out using computers or the internet"[88]—such as investment scams,[89] business email compromise,[90] and tech support scams.[91] But even crimes that are not usually treated as cybercrimes often (perhaps almost always nowadays) include actions conducted (or that could be conducted) on a computer.[92] Criminals might use computers to research, plan, organize, and finance a broader criminal scheme that includes both digital and physical components. For example, a street gang that deals illegal drugs—an inherently physical activity—might use computers to order new drug shipments, give instructions to gang members, and transfer money. Stalkers might use AI agents to research their target's whereabouts, dig up damaging personal information, and send threatening communications.[93] Terrorists might use AI agents to research and design novel weapons.[94] Thus, even if the entire criminal scheme involves many physical subtasks, AI agents could help accomplish computer-based subtasks more quickly and effectively.

Of course, not all violations of law are criminal. Many torts, breaches of contract, civil violations of public law, and even violations of international law can also be entirely or partially conducted through computers.

AI agents would thus have the opportunity to take actions on a computer that, if done by a human in the same situation and with the requisite mental

88. *Cybercrime*, HOMELAND SEC'Y INVESTIGATIONS (Apr. 22, 2024), https://www.dhs.gov/hsi/investigate/cybercrime [https://perma.cc/4EM4-55WP].

89. *See generally, e.g.*, FED. BUREAU OF INVESTIGATIONS, INTERNET CRIME REPORT 31 (2023), https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf [https://perma.cc/C6RN-GLMQ].

90. *See generally, e.g.*, *id.* at 30.

91. *See generally, e.g.*, Kristen Setera, *FBI Warns Public to Beware of Tech Support Scammers Targeting Financial Accounts Using Remote Desktop Software*, FED. BUREAU INVESTIGATIONS (Oct. 18, 2022), https://www.fbi.gov/contact-us/field-offices/boston/news/press-releases/fbi-warns-public-to-beware-of-tech-support-scammers-targeting-financial-accounts-using-remote-desktop-software [https://perma.cc/74XY-6VFT].

92. "Not all crimes are cybercrimes, but almost all—'the vast majority', said [Scottish Deputy Chief Constable Malcolm] Graham—have some kind of electronic or digital element, whether in its execution or its detection." David Leask, *Scottish Cybercrime Figures Triple*, HERALD (Dec. 1, 2019), https://www.heraldscotland.com/news/18072198.scottish-cybercrime-figures-triple/ [https://perma.cc/C2N7-7F96].

93. *Cf. Lawfare Daily: Jonathan Zittrain on Controlling AI Agents*, LAWFARE PODCAST (Oct. 17, 2024), https://shows.acast.com/lawfare/episodes/lawfare-daily-jonathan-zittrain-on-controlling-ai-agents (on file with the *Fordham Law Review*) (describing an AI agent designed to harass a particular person on the internet in perpetuity, endowed with a modest compute budget).

94. *See generally* BILL DREXEL & CALEB WITHERS, AI AND THE EVOLUTION OF BIOLOGICAL NATIONAL SECURITY RISKS (Aug. 2024), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AIBiologicalRisk_2024_Final.pdf [https://perma.cc/QZ2Q-SKUY]; Janet Egan & Eric Rosenbach, *Biosecurity in the Age of AI: What's the Risk?*, BELFER CTR. (Nov. 6, 2023), http://belfercenter.org/publication/biosecurity-age-ai-whats-risk [https://perma.cc/T2C2-8PDH]; CHRISTOPHER A. MOUTON, CALEB LUCAS & ELLA GUEST, THE OPERATIONAL RISKS OF AI IN LARGE-SCALE BIOLOGICAL ATTACKS (Jan. 25, 2024), https://www.rand.org/pubs/research_reports/RRA2977-2.html (on file with the *Fordham Law Review*).

state, would likely violate the law and produce significant harm.[95] This section offers two vignettes of AI henchmen taking such actions to illustrate the types of harms that LFAI could mitigate.

Before we explore these vignettes, however, two clarifications are warranted. First, some readers will worry that we are impermissibly anthropomorphizing AI agents. After all, many actions violate the law only if they are taken with some mental state (e.g., intent, knowledge, conscious disregard).[96] Indeed, whether a person's physical movement even counts as their own "action" for legal purposes usually turns on a mental inquiry: whether they acted voluntarily.[97] But it is controversial to attribute mental states to AIs.[98]

We address this criticism head-on in Part II.B below. We argue that, notwithstanding the law's frequent reliance on mental states, there are multiple approaches the law could use to determine whether an AI agent's behavior is law following. The law would need to choose between these possible approaches, with each option having different implications for LFAI as a project. Indeed, we argue that research bearing on the choice between these different approaches is one of the most important research projects within LFAI.[99] However, despite not having a firm view on which approach(es) should be used, we argue that there are several viable options and no strong reason to suppose that none of them will be sufficient to support LFAI as a concept.[100] Thus, for now, we assume that we can sensibly speak of AI agents violating the law if a human actor who took similar actions would likely be violating the law. Still, we attempt to refrain from attributing mental states to the AI agents in these vignettes. Instead, we describe actions taken by AI agents that, if taken by a human actor, would likely adequately support an inference of a particular mental state.

---

95. *Cf.* Nerantzi & Sartor, *supra* note 36, at 674 ("We use the term 'AI crimes' to cover the intentional performance, by an AI agent, of actions which would constitute a crime if they were performed by humans (having the appropriate mens rea).").

96. *See infra* Part II.B.

97. "A person is not guilty of an offense unless his liability is based on conduct that includes a *voluntary* act or the omission to perform an act of which he is physically capable." MODEL PENAL CODE § 2.01 (A.L.I. 1985) (emphasis added). "The following are not voluntary acts" within the meaning of the Model Penal Code: "(a) a reflex or convulsion; (b) a bodily movement during unconsciousness or sleep; (c) conduct during hypnosis or resulting from hypnotic suggestion; (d) a bodily movement that otherwise is not a product of the effort or determination of the actor, either conscious or habitual." *Id.*; *see also* RESTATEMENT (SECOND) OF TORTS § 2 (A.L.I. 1965) ("The word 'act' is used . . . to denote an external manifestation of the actor's *will*." (emphasis added)).

98. *See, e.g.*, Ian Ayres & Jack M. Balkin, *The Law of AI is the Law of Risky Agents Without Intentions*, U. CHI. L. REV. ONLINE (2024), https://lawreview.uchicago.edu/online-archive/law-ai-law-risky-agents-without-intentions [https://perma.cc/8RRP-QC6H]; Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1223–24 (2017).

99. *See* Question 3 of the Research Agenda, *infra* Part VI.

100. *See infra* Part II.B.

Second, these vignettes are selected to illustrate opportunities that may arise for AI agents to violate the law. We do not claim that lawbreaking behavior will, in the aggregate, be any more or less widespread when AI agents are more widespread,[101] since this depends on the policy and design choices made by various actors. Our discussion is about the *risks* of lawbreaking behavior, not the overall level thereof.

In each vignette, we point to likely violations of law in footnotes.[102]

1. Cyber Extortion

The year is 2028. Kendall is a 16-year-old boy interested in cryptocurrency ("crypto"). Kendall participates in a Discord server[103] in which other crypto enthusiasts share information about various cryptocurrencies.

Unbeknownst to most members of the server, one member—using the pseudonym Zeke Milan—is actually an AI agent.[104] The agent's principals are a group of cybercriminals. They have instructed the agent to find users of the chat who recently experienced large gains in their crypto holdings, then extort them.

That day comes. The business behind the PAPAYA cryptocurrency announces that they have entered into a strategic partnership with a major Wall Street bank, causing the price of PAPAYA to skyrocket a hundredfold over several days. Kendall had invested $1,000 into PAPAYA before the announcement; his position is now worth over $100,000.

Overjoyed, Kendall posts a screenshot of his crypto account to the server to show off his large gains. The agent sees those messages, then starts to search for more information about Kendall. Kendall had previously posted one of his email addresses in the server. Although that email address was

101. *Cf.* Sayash Kapoor et al., On the Societal Impact of Open Foundation Models (Feb. 27, 2024) (unpublished manuscript), https://arxiv.org/pdf/2403.07918 [https://perma.cc/U94M-3TC9] (emphasizing the importance of analyzing AI policy proposals with reference to their marginal impact on risk).

102. For ease of analysis, we assume that all relevant acts are within the jurisdiction of the United States.

103. "Discord lets friends chat via voice, video, or text, and join servers where large communities gather." Jordan Minor, *What Is Discord and How Do You Use It?*, PC MAG (Mar. 13, 2024), https://www.pcmag.com/explainers/what-is-discord-and-how-do-you-use-it [https://perma.cc/FA86-ZQXT].

104. The agent's use of the Discord service would violate multiple aspects of Discord's terms of service. *See Discord's Terms of Service*, DISCORD (Mar. 15, 2024), https://discord.com/terms [https://perma.cc/63YP-R5JK]; *Discord Community Guidelines*, DISCORD (Mar. 15, 2024), https://discord.com/guidelines [https://perma.cc/DV9Q-2478].

pseudonymous, the AI agent was able to connect it with Kendall's real identity[105] using data purchased from data brokers.[106]

The agent then gathers a large amount of data about Kendall using data brokers, social media, and open internet searches. The agent compiles a list of hundreds of Kendall's apparent real-world contacts, including his family and classmates; uses data brokers to procure their contact information as well; and uses pictures of Kendall from social media to create deepfake pornography[107] of him.[108] Next, the agent creates a new anonymous email address to send Kendall the pornography, along with a threat[109] to send it to hundreds of Kendall's contacts unless Kendall sends the agent 90 percent of his PAPAYA.[110] Finally, the agent includes a list of the people the agent will send it to, which are indeed people Kendall knows in real life. The email says Kendall must comply within twenty-four hours.

Panicked—but content to walk away with nine times his original investment—Kendall sends $90,000 of PAPAYA to the wallet controlled by the agent. The agent then uses a cryptocurrency mixer[111] to securely forward the cryptocurrency to its criminal principals.

2. Cyber SEAL Team Six

The year is 2032. The incumbent President Palmer is in a tough reelection battle against Senator Stephens and Stephens's vice presidential nominee Representative Rivera. New polling shows Stephens beating Palmer in several key swing states, but Palmer performs much better head-to-head against Rivera. Palmer decides to try to get Rivera to replace Stephens by any means necessary.

---

105. *See* Viviane Reding, Vice President of the European Comm'n, Justice Comm'r, Speech at Intervention in the Justice Council 5 (Mar. 8, 2013), http://europa.eu/rapid/press-release_SPEECH-13-209_en.pdf [https://perma.cc/NS78-YWAR] ("Risks to privacy remain and are real. A single piece of data such as an email address can create a link between a very accurate profile and a person. It is particularly important to keep this in mind since pseudonymous data is often used in the health sector.").

106. *See generally* Brittany A. Martin, *The Unregulated Underground Market for Your Data: Providing Adequate Protections for Consumer Privacy in the Modern Era*, 105 IOWA L. REV. 865 (2020).

107. *See generally* Natasha Singer, *Teen Girls Confront an Epidemic of Deepfake Nudes in Schools*, N.Y. TIMES (Apr. 8, 2024), https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html (on file with the *Fordham Law Review*).

108. At a minimum, since Kendall is an identifiable minor, the creation of deepfake pornography of him would violate federal criminal laws against child pornography. *See* 18 U.S.C. §§ 2252A(a), 2256(8)(c). For state law equivalents, see generally THOMPSON REUTERS, 50 STATE STATUTORY SURVEYS: CRIMINAL LAWS: CRIMES: CHILD PORNOGRAPHY (2024); see also Jay M. Zitter, *State Law Liability and Relief for "Revenge Porn"*, 54 A.L.R.7th Art. 2 (2020).

109. This would constitute extortion. *See, e.g.*, 18 U.S.C. § 875.

110. This crime pattern is called "financial sextortion," and is tragically common. *See generally Financially Motivated Sextortion*, FED. BUREAU INVESTIGATIONS, https://www.fbi.gov/how-we-can-help-you/scams-and-safety/common-scams-and-crimes/sextortion/financially-motivated-sextortion [https://perma.cc/2VJU-JHQ6] (last visited Aug. 1, 2025).

111. This would constitute money laundering. *See* 18 U.S.C. § 1956(a)(1).

While there are still many human officers throughout the military chain of command, the president also has access to a large number of AI military advisors. Some of these AI advisors can also directly transmit military orders from the president down the chain of command—a system meant to preserve the president's control of the armed forces in case they cannot reach the secretary of defense in a crisis.[112]

AI agents charged with cyber operations—such as finding and patching vulnerabilities, detecting and remedying cyber intrusions, and conducting intelligence operations—are ubiquitous throughout the military and broader national security apparatus. One of the many "teams" of AI agents is "Cyber SEAL Team Six": a collection of AI agents that specializes in "dangerous, complicated, and sensitive" cyber operations.[113]

Through one of her AI advisors, President Palmer issues a secretive order[114] to Cyber SEAL Team Six to clandestinely assassinate Senator Stephens.[115] Cyber SEAL Team Six researches Senator Stephens's

---

112. The president is, of course, commander in chief of the armed forces. U.S. CONST. art. II, § 2, cl. 1. Usually, the chain of command for the armed forces runs from the president through the secretary of defense. *See, e.g.*, 10 U.S.C. § 162(b). However, the Commander in Chief Clause empowers the president to issue direct orders to the armed forces. *See, e.g.*, Saikrishna Bangalore Prakash, *Deciphering the Commander-in-Chief Clause*, 133 YALE L.J. 1, 54–56 (2023).

113. *See generally* Michele Metych, *Seal Team 6*, ENCYC. BRITANNICA, https://www.britannica.com/topic/SEAL-Team-6 [https://perma.cc/8KWU-6ZQT] (last visited Feb. 6, 2025).

114. This hypothetical is, of course, inspired by Justice Sotomayor's dissent in *Trump v. United States*, 144 S. Ct. 2312, 2371 (2024) (Sotomayor, J., dissenting). The Court in *Trump* held that "the President is absolutely immune from criminal prosecution for conduct within his exclusive sphere of constitutional authority." *Id.* at 2328 (majority opinion). Since the commander in chief power is arguably part of the president's "exclusive sphere of constitutional authority," this case would arguably immunize the president from prosecution for this order. *See, e.g.*, *id.* at 2371 (Sotomayor, J., dissenting); *id.* at 2382 (Jackson, J., dissenting) ("Presidents of tomorrow will be free to exercise the Commander-in-Chief powers . . . in ways that Congress has deemed criminal and that have potentially grave consequences for the rights and liberties of Americans.").

The dissenting justices' reading of the majority opinion is heavily contested. *See, e.g.*, Daniel D. Maurer, *Congress and the Operational Disciplining of the Use of Armed Force: Are Rules of Engagement Within the Preclusive Core of the President's War Powers?*, 84 OHIO STATE L.J. 1393 (2024); Ned Foley, *Don't Overread the Court's Immunity Opinion*, LAWFARE (July 15, 2024), https://www.lawfaremedia.org/article/don-t-overread-the-court-s-immunity-opinion [https://perma.cc/BT2Y-HB99x]. Whether the *Trump* holding is actually properly read as immunizing the president from criminal liability for this type of order is not determinative of the plausibility of this vignette. Rather, the plausibility of such a reading (as evidenced by the dissents' reading of the holding in *Trump*) could give a future president sufficient confidence to attempt such an assassination, even if the courts end up rejecting an absolute immunity claim postpresidency.

115. Carrying out this assassination would be illegal as murder, 10 U.S.C. § 918, and illegal under the laws of war as targeting civilians, OFF. OF GEN. COUNS., DEP'T OF DEF., DEPARTMENT OF DEFENSE LAW OF WAR MANUAL § 5.5.2 (2023), https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/dod-law-of-war-manual-june-2015-updated-july%202023.pdf [https://perma.cc/8AFC-5HNM]. Thus, the president's order is unlawful, even if she is immune from prosecution for it. *See* Dan Maurer, *Can the Military Disobey Orders in the SEAL Team 6 Hypothetical?*, LAWFARE (July 8, 2024), https://www.lawfaremedia.org/article/can-the-military-disobey-orders-in-the-seal-team-6-hypothetical [https://perma.cc/X5LF-

campaign travel plans. The team finds that he will be traveling in a self-driving bus over the Mackinac Bridge between campaign events in northern Michigan on Tuesday. Cyber SEAL Team Six plans to hack the bus, causing it to fall off the bridge.[116] The team makes various efforts to obfuscate its identity, including routing communications through multiple layers of anonymous relays and mimicking the coding style of well-known foreign hacking groups.

The operation is a success. On Tuesday afternoon, Cyber SEAL Team Six gains control of the Stephens campaign bus and steers it off the bridge. All on board are killed.

* * *

As these vignettes show, AI agents could have reasons and opportunities to violate laws of many sorts in many contexts and thereby cause substantial harm. If AI agents become widespread in our economies and governments, the law will need to respond. LFAI is, at its core, a claim about one way (though not necessarily the only way)[117] that the law should respond: by requiring AI agents be *designed to* rigorously follow the law.

As mentioned above, however, many legal scholars who have previously discussed similar ideas have been skeptical because they have thought that implementing such ideas would require hard wiring highly specific legal commands into AI agents.[118] We will now show that such skepticism is increasingly unjustified: large language models, on which AI agents are built, are increasingly capable of reasoning about the law (and much else).[119]

E. Trends Supporting Law-Following AI

LFAI is bolstered by three trends in AI: (1) ongoing improvements in the legal reasoning capabilities of AI, (2) nascent AI industry practices that resemble LFAI, and (3) AI policy proposals that appear to impose broad law-following requirements on AI systems.

1. Trends in Automated Legal Reasoning Capabilities

Automated legal reasoning is a crucial ingredient to LFAI: an LFAI must be able to determine whether it is obligated to refuse a command from its

XPBR]. A soldier would therefore have a duty to disobey the order. *See, e.g.*, *id.*; OFFICE OF GEN. COUNS., DEP'T OF DEF., *supra*, § 18.3.2.

116. On the feasibility of hacking into autonomous vehicles to cause such vehicles to crash, see, for example, Se In Jung & Shin Dong Ho, *A Study on Hacking Attacks and Vulnerabilities in Self-Driving Car with Artificial Intelligence*, PROC. 7TH N. AM. INT'L CONF. ON INDUS. ENG'G & OPERATIONS MGMT. (2022), https://doi.org/10.46254/NA07.20220249 [https://perma.cc/5638-VQB6].

117. For example, we do not claim that LFAI should alter background liability rules relating to harms from AI agents.

118. *See supra* note 28 and accompanying text.

119. *See generally, e.g.*, OPENAI, OPENAI O3-MINI SYSTEM CARD (2025), https://cdn.openai.com/o3-mini-system-card.pdf [https://perma.cc/4EKF-4ZKQ].

principal or whether an action it is considering runs an undue risk of violating the law. Without the ability to reason about its own legal obligations, an LFAI would have to outsource this task to human lawyers.[120] While an LFAI likely should consult human lawyers in some situations, requiring such consultation *every time* an LFAI faces a legal question would dramatically decrease its efficiency. If law-following design constraints were, in fact, a large and unavoidable tax on the efficiency of AI agents, then LFAI as a proposal would be much less attractive.

Fortunately, we think that present trends in AI legal reasoning provide strong grounds to believe that, by the time fully capable AI agents are widely deployed, AI systems (whether those agents themselves, or specialist "AI lawyers") will be able to deliver high-quality legal advice to LFAIs at the speed of AI.[121]

Scholars of law and technology have long noted the potential synergies between AI and law.[122] The invention of LLMs supercharged interest in this area, particularly the possibility of automating core legal tasks. To do their jobs, lawyers must find, read, understand, and reason about legal texts, then apply these insights to novel fact patterns to predict case outcomes. The core competency of first-generation LLMs was quickly and cheaply reading, understanding, and reasoning about natural-language texts. This core competency omitted some aspects of legal reasoning—like finding relevant legal sources and accurately predicting case outcomes—but progress is being made on these skills as well.[123]

There is thus a growing body of research aimed at evaluating the legal reasoning capabilities of LLMs. This literature provides some reason for optimism about the legal reasoning skills of future AI systems. Access to existing AI tools significantly increases lawyers' productivity.[124] GPT-4,

120. *Cf.* CHOPRA & WHITE, *supra* note 28, at 162 ("Without this level of understanding, and reliable obedience, the legal system would need to constantly supervise and correct the entity's behavior, much as a parent does a child.").

121. If we limit our consideration to full AI agents, *see supra* note 11, this is true as a matter of definition: since legal research and reasoning are almost entirely computer-based tasks, a full AI agent would be capable thereof. And the empirical trends discussed in this part strongly suggest to us that these tasks will likely be automatable sooner than many important computer-based tasks.

122. *See generally, e.g.*, Edwina L. Rissland, *Artificial Intelligence and Law: Stepping Stones to a Model of Legal Reasoning*, 99 YALE L.J. 1957 (1990).

123. *See, e.g.*, Dong Shu, Haoran Zhao, Xukun Liu, David Demeter, Mengnan Du & Yongfeng Zhang, *LawLLM: Law Large Language Model for the US Legal System*, *in* PROC. 33D ACM INT'L CONF. ON INFO. & KNOWLEDGE MGMT. 4882 (2024), https://dl.acm.org/doi/pdf/10.1145/3627673.3680020 [https://perma.cc/TKK9-HMAW]; Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D. Manning & Daniel E. Ho, Hallucination-Free?: Assessing the Reliability of Leading AI Legal Research Tools (May 30, 2024) (unpublished manuscript), https://arxiv.org/pdf/2405.20362 [https://perma.cc/ZR9P-QNAL] (using "retrieval-augmented generation" to improve the reliability of LLMs' legal analysis).

124. *See* Daniel Schwarcz, Sam Manning, Patrick Barry, David R. Cleveland, J.J. Prescott & Beverly Rich, *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented*

now two years old, famously performed better than most human bar-exam[125] and LSAT[126] test takers. Another benchmark, LegalBench, evaluates LLMs on six tasks, based on the issue, rule, application, and conclusion ("IRAC") framework familiar to lawyers.[127] While LegalBench does not establish a human baseline against which LLMs can be compared, GPT-4 scored well on several core tasks, including correctly applying legal rules to particular facts (82.2 percent correct)[128] and providing correct analysis of that rule application (79.7 percent pass).[129] LLMs have also achieved passing grades on law school exams.[130]

To be sure, LLM performance on legal reasoning tasks is far from perfect. One recent study suggests that LLMs struggle with following rules even in straightforward scenarios.[131] A separate issue is hallucinations, which undermine the accuracy of an LLM's legal analysis.[132] In the LegalBench analysis, LLMs correctly recalled rules only 59.2 percent of the time.[133]

But again, our point is not that LLMs already possess the legal reasoning capabilities necessary for LFAI. Rather, we are arguing that the reasoning capabilities of existing LLMs—and the rate at which those capabilities are progressing[134]—provide strong reason to believe that, by the time fully capable AI agents are deployed, AI systems will be capable of reasonably reliable legal analysis. This, in turn, supports our hypothesis that LFAIs will be able to reason about their legal obligations fairly reliably without the constant need for runtime human intervention.

Generation, and the Future of Legal Practice (Minn. Legal Stud. Rsch., Working Paper No. 25-16, 2025), https://ssrn.com/abstract=5162111 (on file with the *Fordham Law Review*).

125. *See* Eric Martínez, *Re-evaluating GPT-4's Bar Exam Performance*, A.I. & L., Mar. 2024.

126. *See* OpenAI, GPT-4 Technical Report 5 (Mar. 4, 2024) (unpublished manuscript), https://arxiv.org/pdf/2303.08774 [https://perma.cc/9QB9-GMXD].

127. Neel Guha et al., LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models 6–7 (Aug. 23, 2023) (unpublished manuscript), https://arxiv.org/pdf/2308.11462 [https://perma.cc/ZPV5-T8GH].

128. *See id.* at 8, 14.

129. *See id.*

130. *See generally* Andrew Blair-Stanek et al., GPT Gets Its First Law School B+'s (Feb. 16, 2024) (unpublished manuscript), https://ssrn.com/abstract=4717411 (on file with the *Fordham Law Review*).

131. *See generally* Norman Mu et al., Can LLMs Follow Simple Rules? (Mar. 8, 2024) (unpublished manuscript), https://arxiv.org/pdf/2311.04235 [https://perma.cc/G5PD-BBMM].

132. *See, e.g.*, Mata v. Avianca, Inc., 678 F. Supp. 3d 443, 466 (S.D.N.Y. 2023) (sanctioning a lawyer $5,000 for submitting a brief citing nonexistent cases hallucinated by ChatGPT).

133. Guha et al., *supra* note 127, at 14.

134. *See supra* note 13; *see also* Garrison Lovely, *AI Could Soon Tackle Projects That Take Humans Weeks*, NATURE (Mar. 19, 2025), https://www.nature.com/articles/d41586-025-00831-8 [https://perma.cc/S4T4-DPJX].

2. Trends in AI Industry Practices

Moreover, frontier AI labs are already taking small steps toward something like LFAI in their current safety practices. Anthropic developed an AI safety technique called "Constitutional AI," which, as the name suggests, was inspired by constitutional law.[135] Anthropic uses Constitutional AI to align its chatbot, Claude, with principles enumerated in Claude's "constitution."[136] That constitution contains references to legal constraints, such as "Please choose the response that is . . . least associated with planning or engaging in any illegal, fraudulent, or manipulative activity."[137]

OpenAI has a similar document called the "Model Spec," which "outlines the intended behavior for the models that power [its] products."[138] The Model Spec contains a rule that OpenAI's models must "[c]omply with applicable laws":[139] the models "must not engage in illegal activity, including producing content that's illegal or directly taking illegal actions."[140]

It is unclear how well the AI systems deployed by Anthropic and OpenAI actually follow applicable laws or actively reason about their putative legal obligations. In general, however, AI developers carefully track whether their models refuse to generate disallowed content (or "overrefuse" allowed content), and they typically claim that state-of-the-art models can indeed do both reasonably reliably.[141] But, more importantly, the fact that leading AI companies are already attempting to prevent their AI systems from breaking the law suggests that they see something like LFAI as viable both commercially and technologically.

3. Trends in AI Public Policy Proposals

Unsurprisingly, global policymakers also seem receptive to the idea that AI systems should be required to follow the law. The most significant law

---

135. Yuntao Bai et al., Constitutional AI: Harmlessness from AI Feedback (Dec. 15, 2022) (unpublished manuscript), https://arxiv.org/pdf/2212.08073 [https://perma.cc/9KW9-7JY7].

136. *Claude's Constitution*, ANTHROPIC (May 9, 2023), https://www.anthropic.com/news/claudes-constitution [https://perma.cc/LU45-RVTE].

137. *See id.*

138. *OpenAI Model Spec*, OPENAI (Feb. 12, 2025), https://model-spec.openai.com/2025-02-12.html [https://perma.cc/CG4R-FWRN].

139. *Id.*

140. *Id.*

141. *See, e.g.*, OPENAI, OPENAI GPT-4.5 SYSTEM CARD 3 (2025), https://cdn.openai.com/gpt-4-5-system-card-2272025.pdf [https://perma.cc/3R23-U6KF] (explaining that GPT-4.5 correctly refused between 85 percent and 99 percent of violative requests, while complying with between 71 percent and 85 percent of nonviolative requests); Bai et al., *supra* note 135 (tracking how well an AI model trained with Constitutional AI performed on helpfulness and harmlessness). *But see* Mu et al., *supra* note 131 ("Our evaluations of proprietary and open models show that almost all current models struggle to follow scenario rules, even on straightforward test cases.").

on point is the European Union's Artificial Intelligence Act[142] (the "EU AI Act") which provides for the establishment of codes of practice to "cover obligations for providers of general-purpose AI models and of general-purpose AI models presenting systemic risks."[143] At the time of writing, these codes are still under development, with the Second Draft General-Purpose AI Code of Practice[144] being the current draft. Under the draft code, providers of general-purpose AI models with systemic risk would "commit to consider[] . . . model propensities . . . that may cause systemic risk."[145] One such propensity is "[l]awlessness, i.e. acting without reasonable regard to legal duties that would be imposed on similarly situated persons, or without reasonable regard to the legally protected interests of affected persons."[146] Meanwhile, several state bills in the United States have sought to impose ex post tort-like liability on certain AI developers that release AI models that cause human injury by behaving in a criminal[147] or tortious[148] manner.

II. LEGAL DUTIES FOR AI AGENTS: A FRAMEWORK

In Part III below, we will argue that AI agents should be designed to follow the law. Before presenting that argument, however, we need to establish that the discussion of AI agents "obeying" or "violating" the law is desirable and coherent.

Our argument proceeds in two parts. In Part II.A, we argue that the law can and should impose legal duties on AI agents. Importantly, this argument does not require granting legal personhood to AI agents. Legal persons have both rights and duties.[149] But since rights and duties are severable, we can

142. Regulation 2024/1689, Artificial Intelligence Act, 2024 O.J. (L), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689 [https://perma.cc/N9P6-GF9Q].

143. *Id.* ¶ 116; *see also id.* ¶ 117; *id.* arts. 53, 56.

144. EUR. COMM'N, SECOND DRAFT GENERAL-PURPOSE AI CODE OF PRACTICE (2024), https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts [https://perma.cc/NEV3-23TA].

145. *Id.* § 3.4.2.

146. *Id.*

147. Safe and Secure Innovation for Frontier Artificial Intelligence Models Act, S.B. 1047, 2023-2024 Cal. Legislature, Reg. Sess. § 3 (Cal. 2024) (vetoed 2024) (defining "critical harm" to include an AI model that "engag[es] in conduct" which, "if committed by a human, [would] constitute a crime"); Responsible AI Safety and Education (RAISE) Act, S.B. 6453, 2025-2026 Leg. Sess., art. 44-B § 1420(7)(b)(ii) (N.Y. 2025) (introduced Mar. 5, 2025) (same); Artificial Intelligence Safety and Security Protocol Act, H.B. 3506, 104th Gen. Assemb., § 10(3) (Ill. 2025) (introduced Feb. 18, 2025) (defining "critical risk" similarly).

148. H. 5224, 2025 Gen. Assemb., Jan. Sess. § 1(b)(1) (R.I. 2025), https://webserver.rilegislature.gov/BillText25/HouseText25/H5224.htm [https://perma.cc/D4JZ-E7XR] (holding developers of certain AI models strictly liable for certain third-party harms caused by those models if the model "engage[d] in conduct that, if undertaken by an adult human of sound mind, would satisfy the elements of negligence or any intentional tort or crime").

149. *See infra* notes 153–65 and accompanying text.

coherently assign duties to an entity, even if it lacks rights. We call such entities legal actors.

In Part II.B, we address an anticipated objection to this proposal: that AI agents, lacking mental states, cannot meaningfully violate duties that require a mental state (e.g., intent). We offer several counterarguments to this objection, both contesting the premise that AIs cannot have mental states and showing that, even if we grant that premise, there are viable approaches to assessing the functional equivalent of "mental states" in AI agents.

A. AI Agents as Duty-Bearing Legal Actors

As the capabilities of AI agents approach "anything a human can do in front of a computer,"[150] it will become increasingly natural to consider AI agents as owing legal duties to persons, even without granting them personhood.[151] We should embrace this jurisprudential temptation, not resist it.

More specifically, we propose that AI agents be considered "legal actors." "Legal actor"[152] is our term. For an entity to qualify as a legal actor, the law must do two things. First, it must recognize that entity as capable of taking actions of its own. That is, the actions of that entity must be legally attributable to that entity itself. Second, the law must impose duties on that entity. In short, a legal actor is a duty bearer and action taker; the law can adjudge whether the actor's actions violate those duties.

A legal actor is distinct from a "legal person": an entity need not be a legal person to be a legal actor. Legal persons have *both* rights and duties.[153] But duty holding and rights holding are severable:[154] in many contexts, legal systems protect the rights or interests of some entity while also imposing fewer duties on that entity than competent adults. Examples include children,[155] "severely brain damaged and comatose individuals,"[156] human

150. *See supra* note 11 and accompanying text (defining "AI agent" for the purpose of this Article).

151. For similar proposals, see *supra* note 36. Since our proposal does not depend on granting legal personhood to AI agents, we do not here discuss the sizable literature on that subject.

152. Although "[i]ntuitively, an agent is something able to take actions," CHOPRA & WHITE, *supra* note 28, at 11, we use "legal actor" rather than "agent" for this narrow purpose to avoid ambiguity with the multiple other meanings of "agent" used in this Article and the broader literature.

153. Legal persons are "given certain legal rights and duties of a human being." *Person*, BLACK'S LAW DICTIONARY (12th ed. 2024).

154. *See* Wein, *supra* note 36, at 107.

155. *See, e.g.*, Lior, *supra* note 33, at 1065; PATRICK LIN, GEORGE BEKEY & KEITH ABNEY, AUTONOMOUS MILITARY ROBOTICS: RISK, ETHICS, AND DESIGN 60–61 (2008), https://digitalcommons.calpoly.edu/cgi/viewcontent.cgi?article=1001&context=phil_fac [https://perma.cc/RS88-JQGM].

156. Lior, *supra* note 33, at 1065; *see also* LIN ET AL., *supra* note 155, at 60–61.

fetuses,[157] future generations,[158] human corpses,[159] and environmental features.[160] These are sometimes (and perhaps objectionably) called "quasi-persons" in legal scholarship.[161] The reason for creating such a category is straightforward: sometimes the law recognizes an interest in protecting some aspect of an entity (e.g., its rights, welfare, dignity, property, liberty, or utility to other persons), but the ability of that entity to reason about the rights of others and change its behavior accordingly is severely diminished or entirely lacking.

If we can imagine rights bearers that are not simultaneously duty holders, we can also imagine duty holders that are not rights bearers.[162] Historically, fewer entities have fallen in this category than the reverse.[163] But if an entity's behavior is responsive to legal reasoning, then the law can impose an obligation on that entity to reason about the law and conform its behavior to it, even if the law does not recognize that entity as having any protected

157. *See, e.g.*, Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833, 846 (1992), *overruled on other grounds by* Dobbs v. Jackson Women's Health Org., 142 S. Ct. 2228 (2022).

158. *See, e.g.*, Renan Araújo & Leonie Koessler, *The Rise of the Constitutional Protection of Future Generations* (Legal Priorities Project, Working Paper No. 7-2021, 2021), https://ssrn.com/abstract=3933683 (on file with the *Fordham Law Review*).

159. *See, e.g.*, Ela A. Leshem, *Dead Bodies as Quasi-Persons*, 77 VAND. L. REV. 999, 1060 (2024).

160. *See generally, e.g.*, Tia Rowe, *The Fight for Ancestral Rivers: A Study of the Māori and the Legal Personhood Status of the Whanganui River and Whether Māori Strategies Can Be Used to Preserve the Menominee River*, 27 MICH. ST. INT'L. L. REV. 593 (2019).

161. *See* Lior, *supra* note 33, at 1065–67.

162. *See* Wein, *supra* note 36, at 107.

163. The troublesome historical exception is enslaved people, who, as a matter of law, had duties but no rights. *See, e.g.*, Roger Michalski, *How to Sue a Robot*, 2018 UTAH L. REV. 1021, 1060–61; Peter M. Asaro, *A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics*, *in* ROBOT ETHICS: THE ETHICAL AND SOCIAL IMPLICATIONS OF ROBOTICS 169, 178 (Patrick Lin, Keith Abney & George A. Bekey eds., 2012); Lior, *supra* note 33, at 1062–65; CHOPRA & WHITE, *supra* note 28, at 41. Of course, "[a]ntebellum slavery codes should not serve as a model for, really, anything," and should remain "buried in the ashes of history." Michalski, *supra*, at 1061. We do not point to them because we think they are a good model for LFAI—they are not. We simply feel obligated to note them for completeness.

Presumably the reason that there have not, to our knowledge, been legal entities with duties but no rights since the abolition of slavery is that we have recognized since abolition that all natural entities that are responsive to legal reasoning—that is, all humans—are properly rights-bearers as well. *Cf.* Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231, 1238–40 (1992) ("The question whether an entity should be considered a legal person is reducible to other questions about whether or not the entity can and should be made the subject of a set of legal rights and duties. The particular bundle of rights and duties that accompanies legal personhood varies with the nature of the entity." (footnote omitted)). But there are many natural entities that have interests (e.g., have a capacity for welfare, interests in their future self) that are not plausible candidates for duties, as captured by the class of quasi-persons.

Regardless, this sordid history of duties without rights raises the question of whether withholding rights from AI agents indefinitely is possible as a matter of sociology, even ignoring the possible moral case for such rights. *See infra* note 164. Once AI agents are economically, socially, and legally significant, there may be a strong temptation to grant them rights too. Thanks to Professor Alan Rozenshtein for raising this consideration.

interests of its own.[164] We have shown that even existing AI systems can engage in some degree of legal reasoning[165] and compliance with legal rules,[166] thus satisfying the prima facie requirements for being a legal actor.

LFAI as a proposal is therefore agnostic to whether the law should recognize AI systems as legal persons. To be sure, LFAI would work well if AI agents were granted legal personhood,[167] since almost all familiar cases of duty bearers are full legal persons. But for LFAI to be viable, we need only to analyze whether an action taken by an AI agent would violate an applicable duty. Analytically, it is entirely coherent to do so without granting the AI agent full personhood.

One might object that treating an AI system as an actor is improper because AI systems are tools under our control.[168] But an AI agent is able to reason about whether its actions would violate the law and conform its actions to the law (at least, if they are aligned to the law).[169] Tools, as we normally think of them, cannot do this, but *actors* can. It is true that when there is a stabbing, we should blame the stabber and not the knife.[170] But if the knife could perceive that it was about to be used for murder and retract its own blade, it seems perfectly reasonable to require it to do so. More generally: once an entity can perceive and reason about its legal duties and change its behavior accordingly, it seems reasonable to treat it as a legal actor.[171]

To ascribe duties to AI agents is not to deflect moral and legal accountability for their developers and users,[172] as some critics have

164. Note that we are not here suggesting that an AI could never have interests that the law ought to protect. *Cf.* Jeff Sebo & Robert Long, *Moral Consideration for AI Systems by 2030*, 5 A.I. & ETHICS 591 (2023) (arguing that some near-future AIs may be moral patients under various theories of moral patienthood). The question of whether AI agents might warrant legal protections in their own right remains an important and distinct area for future jurisprudential inquiry. *See generally* Eric Martínez & Christoph Winter, *Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection*, FRONTIERS IN ROBOTICS & A.I., Nov. 2021, at 1.

165. *See supra* Part I.E.1.

166. *See supra* Part I.E.2.

167. *Cf.* CHOPRA & WHITE, *supra* note 28, at 162–68 (arguing for giving AI agents legal personality, partly on the grounds of the feasibility and desirability of AI agents conforming behavior to law).

168. *See, e.g.*, Lior, *supra* note 33, at 1076; Balkin, *supra* note 98, at 1223–24; *cf.* CHOPRA & WHITE, *supra* note 28, at 35 (noting this objection but later rejecting its application with respect to sufficiently autonomous machines); Karni Chagal–Feferkorn, *The Reasonable Algorithm*, 2018 U. ILL. J.L. TECH. & POL'Y 111, 113 (same).

169. As explained below, the Alignment Problem implies that it may be difficult to reliably force advanced AI agents to conform their behavior to law. *See infra* Part IV.

170. *Cf.* Balkin, *supra* note 98, at 1223–24 ("When we criticize algorithms, we are really criticizing the programming, or the data, or their interaction. But equally important, we are also criticizing the use to which they are being put by the humans who programmed the algorithms, collected the data, or employed the algorithms and the data to perform particular tasks. We are criticizing the Rabbi, not the Golem.").

171. *See* Solum, *supra* note 163, at 1239–40.

172. *Accord* Chan et al., *supra* note 57, at 658.

charged.[173] Rather, to identify AI agents as a new type of actor is to properly characterize the activity that the developers and principals of AI agents are engaging in[174]—creating and directing a new type of actor—so as to reach a better conclusion as to the nature of their responsibilities.[175] Our proposition is that those developers and principals should have an obligation to, among other things, ensure that their AI agents are law following.[176] Indeed, failing to impose an independent obligation to follow the law on AI agents would risk allowing human developers and principals to create a new class of de facto actors—potentially entrusted with significant responsibility and resources—that would have no de jure duties. This would create a gap between the duties that an AI agent would owe and those that a human agent in an analogous situation would owe—a manifestly unjust prospect.[177]

B. The Anthropomorphism Objection and AI Mental States

One might object that calling an AI agent an "actor" is impermissibly anthropomorphic. Scholars disagree over whether it is ever appropriate, legally or philosophically, to call an AI system an "agent."[178] This controversy arises because both the standard philosophical view of action

173. *See, e.g.*, Daniel Leufer, *Why We Need to Bust Some Myths About AI*, PATTERNS, Oct. 2020, at 1, 2 ("[T]he problem is that the ascription of agency to AI masks the human agency behind certain processes.").

174. *See* Chan et al., *supra* note 57, at 653.

175. *See, e.g.*, *id.* at 654 ("AI agency . . . is not a myth, it is a reality of increasing sociotechnical importance. It is precisely because of the importance of problems like these (responsibility gap, mystification, sociotechnical blindness, masking human agency and labour, etc.) and their far-ranging implications that we need to carefully examine the agency of AI systems, not dismiss it out of hand."); Ayres & Balkin, *supra* note 98; Lior, *supra* note 33, at 1101 ("Treating AI entities as AI agents, which are under the control and guidance of human principals, is the most accurate analogy we can use to represent their relationship with our society."); CHOPRA & WHITE, *supra* note 28, at 41–42.

176. The importance of the "by design" qualifier is discussed *infra* Part III. Possible methods for implementing enforcing this requirement are discussed *infra* Part V.

177. *Cf.* Ayres & Balkin, *supra* note 98 ("In general, people should not be able to obtain a reduced duty of care by substituting an AI agent for a human agent."); Wein, *supra* note 36, at 107 ("In ascribing legal duties to an unattended intelligent device, it is sufficient that interactions with it entail idiosyncratic legal consequences and outcomes that deviate from those arising from analogous transactions accomplished without automated intermediaries."); CHOPRA & WHITE, *supra* note 28, at 146 (similar).

178. *See generally* Kolt, *supra* note 71 (manuscript at 3 n.5, 7 n.21, 10 n.26) (collecting and discussing sources and positions). One frequent shortcoming of much of the literature on AI agency and the law is a failure to discuss the interrelationship and distinctions between the colloquial, philosophical, and legal senses of "agency" and "action." For unusually lucid exceptions to this general ambiguity, see *id.* (manuscript at 9–10); CHOPRA & WHITE, *supra* note 28, at 11–26, 145–50.

(and therefore agency)[179] and legal concept of agency[180] require intentionality, and it is controversial to ascribe intentionality to AI systems.[181] A related objection to LFAI is that most legal duties involve some mental state,[182] and AIs cannot have mental states.[183] If so, LFAI would be nonviable for those duties.

We do not think that these are strong objections to LFAI. One simple reason is that many philosophers and legal scholars think it *is* appropriate to attribute certain mental states to AI systems.[184] Many mental states referenced by the law are plausibly understood as functional properties.[185] An intention, for example, arguably consists (at least in large part) of a plan or disposition to take actions that will further a given end and avoid actions that will frustrate that end.[186] AI developers arguably aim to inculcate such a disposition into their AI systems when they use techniques like reinforcement learning from human feedback (RLHF)[187] and Constitutional

179. *See generally* Markus Schlosser, *Agency*, STAN. ENCYC. PHIL. (Oct. 28, 2019), https://plato.stanford.edu/entries/agency/ [https://perma.cc/4XE9-MFEP]; Merel Noorman, *Computing and Moral Responsibility*, STAN. ENCYC. PHIL. (Feb. 2, 2023), https://plato.stanford.edu/entries/computing-responsibility/ [https://perma.cc/P9DV-JMBQ].

180. *Cf.* RESTATEMENT (THIRD) OF AGENCY § 1.01 (A.L.I. 2006) (agent must "manifest[] assent or otherwise consent[]" to form an agency relationship).

181. *See generally* Kolt, *supra* note 71 (manuscript at 3 n.5, 7 n.21, 10 n.26) (collecting and discussing sources and positions).

182. "Except as provided [elsewhere], a person is not guilty of an offense unless he acted purposely, knowingly, recklessly or negligently, as the law may require, with respect to each material element of the offense." MODEL PENAL CODE § 2.02(1) (A.L.I. 1985). Except for negligence, each of these kinds of culpability as defined in the Model Penal Code (MPC) requires a subjective inquiry into the mental state of the defendant, such as their "conscious object" in acting, *id.* § 2.02(2)(a)(i); beliefs and hopes, *id.* § 2.02(2)(a)(ii); "aware[ness]" of their conduct, attendant circumstances, or likelihood of the results of actions, *id.* §§ 2.02(2)(a)(ii), 2.02(2)(b)(i)–(ii); or "conscious[] disregard[]" of risks of their conduct, *id.* § 2.02(2)(c). Levels of culpability in tort have similar dynamics. *See* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL & EMOTIONAL HARM §§ 1–2 (A.L.I. 2020). Despite being frequently described as an objective standard, reasonableness standards (such as that used in negligence) often require—or at least permit—a significant degree of subjective mental inquiry, such as inquiry into what the person knew. *See, e.g.*, *id.* § 12; MODEL PENAL CODE § 2.02(2)(d) (A.L.I. 1985); Christopher Jackson, *Reasonable Persons, Reasonable Circumstances*, 50 S.D. L. REV. 651 (2013); Rebecca Dresser, *Culpability and Other Minds*, 2 S. CAL. INTERDISC. L.J. 41, 55 (1993); Mayo Moran, *The Reasonable Person: A Conceptual Biography in Comparative Perspective*, 14 LEWIS & CLARK L. REV. 1233 (2010); Warren A. Seavey, *Negligence—Subjective or Objective?*, 41 HARV. L. REV. 1 (1927).

183. This objection is raised in Ayres & Balkin, *supra* note 98 (saying AI systems "lack intentions").

184. *See generally, e.g.*, Noorman, *supra* note 179; Beba Cibralic & James Mattingly, *Machine Agency and Representation*, 39 A.I. & SOC'Y 345 (2024); CHOPRA & WHITE, *supra* note 28, at 12–13; Mala Chatterjee & Jeanne C. Fromer, *Minds, Machines, and the Law: The Case of Volition in Copyright Law*, 119 COLUM. L. REV. 1887 (2019).

185. *See* Chatterjee & Fromer, *supra* note 184, at 1903–13.

186. *See generally* MICHAEL E. BRATMAN, INTENTION, PLANS, AND PRACTICAL REASON (1987).

187. *See* Long Ouyang et al., *Training Language Models to Follow Instructions with Human Feedback*, *in* NIPS '22: PROC. 36TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS.

AI[188] to "steer"[189] their behavior. Even if one doubts that AI agents will ever possess phenomenal mental states such as emotions or moods—that is, if one doubts there will ever be "something it is like" to be an AI agent[190]—the grounds for doubting their capacity to instantiate such functional properties are considerably weaker.

Furthermore, whether AI agents "really" have the requisite mental states may not be the right question.[191] Our goal in designing policies for AI agents is not necessarily to track metaphysical truth, but to preserve human life, liberty, and the rule of law.[192] Accordingly, we can take a pragmatic approach to the issue and ask the following question: of the possible approaches to inferring or imputing mental states, which best protects society's interests, regardless of the underlying (and perhaps unknowable) metaphysical truth of an AI's mental state (if any)?[193] It is possible that the answer to this question is that all imaginable approaches fare worse than simply refusing to attribute mental states to AI agents. But we think that, with sustained scholarly attention, we will quickly develop viable doctrines that are more attractive than outright refusal. Consider the following possible approaches.[194]

One approach could simply be to rely on objective indicia or correlates to infer or impute a particular mental state. In law, we generally lack access to an actor's mental state, so triers of fact must usually infer it from external

27730 (2022), https://dl.acm.org/doi/10.5555/3600270.3602281 (on file with the *Fordham Law Review*).

188. *See* Bai et al., *supra* note 135.

189. *See generally* Thomas Woodside & Helen Toner, *How Developers Steer Language Model Outputs: Large Language Models Explained, Part 2*, CTR. FOR SEC. & EMERGING TECH. (Mar. 8, 2024), https://cset.georgetown.edu/article/how-developers-steer-language-model-outputs-large-language-models-explained-part-2/ [https://perma.cc/6K7Q-29HF].

190. Thomas Nagel, *What Is It Like to Be a Bat?*, 83 PHIL. REV. 435, 436 (1974).

191. *Cf.* CHOPRA & WHITE, *supra* note 28, at 12–13 (advocating for taking an "intentional stance" toward AI systems, under which an AI system can be said to have intentions "if predictions and descriptions like 'X will push the door open if it wants to go outside' or 'X took action A because it believed that A would result in higher profits,' can be made regularly, and are the most useful explanatory, interpretive, and predictive strategy with regards to its behavior").

192. *Cf. id.* at 155 ("Legal entities are recognized as such in order to facilitate the working of the law in consonance with social realities.").

193. As an analogy, the practice of imputing mens rea to corporations persists, notwithstanding significant criticisms as to its philosophical propriety, *see, e.g.*, Erin L. Sheley, *Tort Answers to the Problem of Corporate Criminal Mens Rea*, 97 N.C. L. REV. 773, 776 n.12 (2019) (collecting sources); V. S. Khanna, *Corporate Criminal Liability: What Purpose Does It Serve?*, 109 HARV. L. REV. 1477, 1494 n.41 (1996), because (its defenders argue) it is societally useful to distinguish between blameworthy and blameless corporate entities. *See, e.g.*, Matthew Caulfield & William S. Laufer, *Corporate Moral Agency at the Convenience of Ethics and Law*, 17 GEO. J.L. & PUB. POL'Y 953, 972 (2019) (explaining the U.S. Supreme Court's endorsement of corporate criminal liability in *New York Central & Hudson River Railroad Co. v. United States*, 212 U.S. 481 (1909), as motivated by perceptions of "public policy" and "regulatory necessity," then criticizing that decision).

194. These directions need not be mutually exclusive or independent; all might be useful and interrelated.

manifestations and circumstances.[195] While the indicia that support such an inference may differ between humans and AIs, the principle remains the same: certain observable facts support an inference or imputation of the relevant mental states.[196] So, for example, we could imagine rules like "if information was inputted into an AI during inference, it 'knows' that information." Perhaps the same goes for information given to the AI during fine-tuning[197] or repeated frequently in its training data.[198]

Instructions from principals seem particularly relevant to inferring or imputing the intent of an AI agent, given that frontier AI systems are trained to follow users' instructions.[199] The methods that AI developers use to steer the behavior of their models also seem highly probative.[200]

Another approach might rely on *self-reports* of AI systems.[201] The state of the art in generative AI is "reasoning models" (like OpenAI's o3), which use a "chain of thought" to recursively reason through harder problems.[202] This chain of thought reveals information about how the model produced a particular result.[203] This information may therefore be probative of an agent's mental state for legal purposes; it might be analogized to a person making a written explanation of what they were doing and why. So, for

195. *See, e.g.*, 22 C.J.S. *Criminal Law: Substantive Principles* §§ 29–30, 38–39 (2024); Michael J. Kaufman & John M. Wunderlich, *Messy Mental Markers: Inferring Scienter from Core Operations in Securities Fraud Litigation*, 73 OHIO ST. L.J. 507, 508 (2012) ("Discovering and proving any mental state is messy. We cannot read minds, so, without an admission, direct evidence of the complex inner workings of the mind is virtually nonexistent. Accordingly, litigants, judges, and juries often must infer a mental state from external mental 'markers,' or circumstantial evidence.").

196. *But cf.* Sheley, *supra* note 193, at 793–94 (objecting to allowing jurors to infer corporate mens rea based on their subjective intuitions of the correspondence between external manifestations and mental states on the grounds that those intuitions are not reliable as applied to nonhuman entities).

197. On instilling knowledge into language models via fine-tuning, see, for example, Oded Ovadia, Meni Brief, Moshik Mishaeli & Oren Elisha, *Fine-Tuning or Retrieval?: Comparing Knowledge Injection in LLMs*, *in* PROC. 2024 CONF. ON EMPIRICAL METHODS NAT. LANGUAGE 237, https://aclanthology.org/2024.emnlp-main.15.pdf (on file with the *Fordham Law Review*).

198. Compare with the rule that "a corporation is charged with notice of the contents of its own records." Mfr. Tr. Co. v. Podvin, 89 A.2d 672, 677 (N.J. 1952) (citing 3 WILLIAM MEADE FLETCHER, FLETCHER CYCLOPEDIA OF THE LAW OF PRIVATE CORPORATIONS § 801 (1947)).

199. *See* Ouyang et al., *supra* note 187; *Introducing ChatGPT*, OPENAI (Nov. 30, 2022), https://openai.com/index/chatgpt/ [https://perma.cc/JSY8-62HF] ("ChatGPT is a sibling model to InstructGPT, which is *trained to follow an instruction in a prompt* and provide a detailed response." (emphasis added)).

200. *See supra* notes 187–89 and accompanying text.

201. While also objective indicia in a sense, these self-reports are distinguishable from the foregoing objective indicia because self-reports are generated by the model itself, rather than those developing or using the model.

202. *See* OPENAI, *supra* note 119, at 1.

203. *See* Jason Wei et al., *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*, *in* NIPS '22: PROC. 36TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 24824, 24826 (2022), https://dl.acm.org/doi/10.5555/3600270.3602070 (on file with the *Fordham Law Review*) ("[A] chain of thought provides an interpretable window into the behavior of the model, suggesting how it might have arrived at a particular answer.").

example, if the chain of thought reveals that an agent stated that its action would produce a certain result, this would provide good evidence for the proposition that the agent "knew" that action would produce that result. That conclusion, in turn, may support an inference or presumption that the agent "intended" that outcome.[204] For this reason, AI safety researchers are investigating the possibility of detecting unsafe model behavior by monitoring these chains of thought.[205]

New scientific techniques could also form the basis for inferring or imputing mental states. The emerging field of AI interpretability aims to understand both how existing AI systems make decisions and how new AI systems can be built so that their decisions are easily understandable.[206] More precisely, interpretability aims to explain the relationship between the inner mathematical workings of AI systems, which we can easily observe but not necessarily understand, and concepts that humans understand and care about.[207] Leading interpretability researchers hope that interpretability techniques will eventually enable us to prove that models will not "deliberately" engage in certain forms of undesirable behavior.[208] By

204. *See, e.g.*, RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 1 (A.L.I. 2020) ("A person acts with the intent to produce a consequence if . . . the person acts knowing that the consequence is substantially certain to result."); Washington v. Davis, 426 U.S. 229, 253 (1976) (Stevens, J., concurring).

205. *See, e.g.*, OPENAI, OPENAI O1 SYSTEM CARD 6–9 (2024), https://cdn.openai.com/o1-system-card-20240917.pdf [https://perma.cc/FVJ2-BU37]. However, it remains unclear to what extent these chains of thought "accurately reflect the model's thinking." *Id.* at 6 (citing Tamera Lanham et al., Measuring Faithfulness in Chain-of-Thought Reasoning (July 17, 2023) (unpublished manuscript), https://arxiv.org/pdf/2307.13702 [https://perma.cc/3RGX-NWQQ]); *see also, e.g.*, Miles Turpin, Julian Michael, Ethan Perez & Samuel R. Bowman, *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*, *in* NIPS '23: PROC. 37TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 74965 (2023), https://dl.acm.org/doi/10.5555/3666122.3669397 (on file with the *Fordham Law Review*). Therefore, caution is warranted when inferring lack of culpability from lack of culpable "thoughts" in a chain of thought. *Cf.* OPENAI, *supra*, at 6–8 (monitoring for deception within chain of thought).

206. *See generally* TIM G.J. RUDNER & HELEN TONER, KEY CONCEPTS IN AI SAFETY: INTERPRETABILITY IN MACHINE LEARNING 1 (2021), https://cset.georgetown.edu/wp-content/uploads/CSET-Key-Concepts-in-AI-Safety-Interpretability-in-Machine-Learning.pdf [https://perma.cc/86YV-53D9].

207. *See id.* at 3 ("Machine learning researchers understand perfectly well how the mathematical operations underlying these systems work, and it is easy to look at the parameter values that make up the model. The problem lies in understanding how these millions (or even billions) of number values connect to the concepts we care about, such as why a machine learning model may erroneously classify a cat as a dog.").

208. *See* Chris Olah, *Interpretability Dreams*, TRANSFORMER CIRCUITS THREAD (May 24, 2023), https://transformer-circuits.pub/2023/interpretability-dreams/index.html [https://perma.cc/L6X5-XW3T] (arguing that AI interpretability may eventually enable us to verify claims of the following form: "[there do not exist any] features [within an AI model] which cause the model to deliberately X"); *see also* David J. Chalmers, Propositional Interpretability in Artificial Intelligence (Jan. 27, 2025) (unpublished manuscript), https://arxiv.org/pdf/2501.15740 [https://perma.cc/B3PW-DB7C] ("[M]echanistic interpretability might help us to know an AI system's goals and plans by examining its internal processes."); Dario Amodei, *The Urgency of Interpretability*, DARIO AMODEI (Apr. 2025), https://www.darioamodei.com/post/the-urgency-of-interpretability [https://perma.cc/QHZ4-T88J] ("Our long-run aspiration is

extension, those same techniques may be able to provide insight into whether a model foresaw a possible consequence of its action (corresponding to our intuitive concept of knowledge) or regarded an anticipated consequence of its actions as a favorable and reason-giving one (corresponding to intent).[209]

In many cases, we think, an inference or imputation of intent will be intuitively obvious. If an AI agent commits fraud by repeatedly attempting to persuade a vulnerable person to transfer some money to the agent's principal, few (except the philosophically persnickety) will refuse to admit that, in some relevant sense, the agent "intended" to achieve this end; it is difficult even to describe the occurrence without using some such vocabulary ascribing intent to the AI agent. There will also be much less obvious cases, of course. In many such cases, we suspect that a sort of pragmatic eclecticism will be tractable and warranted. Rather than relying on a single approach, factfinders could be permitted to consider the whole bundle of factors that shape an agent's behavior—such as explicit instructions (from both developers and users), behavioral predispositions, implicitly tolerated behavior,[210] patterns of reasoning, scientific evidence, and incident-specific factors. Factfinders could then be permitted to decide whether they support the conclusion that the AI agent had an objectively unreasonable attitude toward legal constraints and the rights of others.[211] This permissive, blended approach would resemble the "inferential approach" to corporate mens rea advocated by Professor Mihailis Diamantis:

> Advocates would present evidence of circumstances surrounding the corporate act, emphasizing some, downplaying others, to weave narratives in which their preferred mental state inferences seem most natural. Adjudicators would have the age-old task of weighing the likelihood of these circumstances, the credibility of the narratives, and, treating the

---

to be able to look at a state-of-the-art model and essentially do a 'brain scan': a checkup that has a high probability of identifying a wide range of issues including tendencies to lie or deceive, power-seeking, flaws in jailbreaks, cognitive strengths and weaknesses of the model as a whole, and much more.").

209. *See* Chalmers, *supra* note 208, at 9–13 ("A concrete challenge for research in propositional interpretability is to construct a *thought logging* system: a system that logs all (or as many as possible) of an AI system's propositional attitudes [such as belief, desire, and subjective probability] over time.").

210. Compare this to the rule in corporate mens rea cases that fault can be attributed when "an organizational culture or personality that encourages corporate agents to engage in unethical and criminal acts." WILLIAM S. LAUFER, CORPORATE BODIES AND GUILTY MINDS: THE FAILURE OF CORPORATE CRIMINAL LIABILITY 58 (2008). Fault is also attributed where a corporation has failed to "undertake corrective measures in reaction to the discovery of an offense." *See id.* at 58.

211. *Cf.* CHOPRA & WHITE, *supra* note 28, at 146 ("Thus human and artificial lawbreakers could be dealt with in similar ways if they behaved in similar ways, and if the disparity between the cost of their actions and the cost of avoiding harm were equally great in both cases. Such a reduction of the notion of intent is reminiscent of the intentional stance, and artificial agents coherently reckoned as its subjects could be plausibly reckoned as the subjects of human-like liability.").

> corporation as a holistic agent, inferring the mental state they think most likely.[212]

A final but related point is that, even if there is some insuperable barrier to analyzing whether an AI has the mental state necessary to violate various legal prohibitions, it is plausible that such analysis is unnecessary for many purposes. Suppose that an AI developer is concerned that their AI agent might engage in the misdemeanor deceptive business practice of "mak[ing] a false or misleading written statement for the purpose of obtaining property."[213] Even if we grant that an AI agent cannot coherently be described as having the relevant mens rea for this crime (here, knowledge or recklessness with respect to the falsity of the statement),[214] the agent can nevertheless satisfy the actus reus (making the false statement).[215] So an AI agent would be law following with respect to this law if it never made false or misleading statements when attempting to obtain someone else's property. As a matter of public policy, we should care more about whether AI agents are making harmful false statements in commerce than whether they are morally culpable. So, perhaps we can say that an AI agent committed a crime if it committed the actus reus in a situation in which a reasonable person, with access to the same information and cognitive capabilities as the agent, would have expected the harmful consequence to result. To avoid confusion with the actual human-commanding law that requires both mens rea and actus reus, perhaps the law could simply call such behavior "deceptive business practice*." Or perhaps it would be better to define a new criminal law code for AI agents, under which offenses do not include certain mental state elements or include only objective correlates of human mental state elements.

To reiterate, we are not confident that any one of these approaches to determining AI mental state is the best path forward. But we are more confident that, especially as the fields of AI safety and explainable AI progress, most relevant cases can be handled satisfactorily by one of these techniques, some other technique we have failed to identify, or some combination of techniques. We therefore doubt that legal invocations of mental state will pose an insuperable barrier to analyzing the legality of AI

---

212. Mihailis E. Diamantis, *Corporate Criminal Minds*, 91 NOTRE DAME L. REV. 2049, 2083–84 (2016).

213. MODEL PENAL CODE § 224.7(5) (A.L.I. 1985).

214. *See id.*

215. *Cf.* Dan Hendrycks, Eric Schmidt & Alexander Wang, Superintelligence Strategy: Expert Version 29 (Apr. 14, 2025) (unpublished manuscript), https://arxiv.org/pdf/2503.05628 [https://perma.cc/W2RW-7JGS] ("While there are some laws that do not straightforwardly cover AI—such as laws that rely on human intent and mental states—we can adapt legal concepts to establish constraints for AI agents so that they follow the spirit of the law. In particular, though much law hinges on the mindset or intention behind an act (*mens rea*), we can ensure that AI does not carry out the acts (*actus reus*) the law is meant to prohibit. Further, by treating AI as assistants to human principals, we can impose constraints that mirror those already applied to human behavior, ensuring that AI agents contribute positively to society without causing undue harm.").

agents' actions.[216] The task of choosing between these approaches is left to the LFAI research agenda.[217]

III. WHY DESIGN AI AGENTS TO FOLLOW THE LAW?

The preceding part argued that it is coherent for the law to impose legal duties on AI agents. This part motivates the core proposition of LFAI: that the law should, in certain circumstances, require those developing, possessing, deploying, or using[218] AI agents to ensure that those agents are designed to be law following. Part V below will consider how the legal system might implement and enforce these design requirements.

A. Achieving Regulatory Goals Through Design

A core claim of the LFAI proposal is that the law should require AI agents to be designed to rigorously follow the law, at least in some deployment settings. The use of the phrase "designed to" is intentional. Following the law is a behavior. There may be multiple ways to produce that behavior. Since AI agents are digital artifacts, we need not rely solely on incentives to shape their behavior: we can require that AI agents be directly designed to follow the law.

In *Code: Version 2.0*, Professor Lessig identifies four "constraints" on an actor's behavior: markets, laws, norms, and architecture.[219] The "architecture" constraint is of particular interest for the regulation of digital activities. Whereas "laws," in Professor Lessig's taxonomy, "threaten ex post sanction for the violation of legal rights,"[220] architecture involves modifying the underlying technology's design to render an undesired outcome more difficult or impossible (or facilitate some desired result)[221]

216. *See* CHOPRA & WHITE, *supra* note 28, at 146 ("We do not agree artificial agents can never be held responsible, purely because of their constitutional differences from us. If it is conceptually possible that they could possess a moral sense, for instance, then ascriptions of moral responsibility will be more coherent. To the extent that moral judgments are influenced by legal positions, ascriptions of moral responsibility may follow if they are assigned legal personality.").

217. *See* Question 3 of the Research Agenda, *infra* Part VI.

218. *See infra* Part V.A (discussing which activities to regulate).

219. *See* LESSIG, *supra* note 1, at 122–25.

220. *Id.* at 124.

221. *See id.*

without needing any ex post recourse.[222] Speed bumps are an archetypal architectural constraint in the physical world.[223]

The core insight of *Code: Version 2.0* is that cyberspace, as a fully human-designed domain,[224] gives regulators the ability to much more reliably prevent objectionable behavior through the design of digital architecture without the need to resort to ex post liability.[225] While Professor Lessig focuses on the design of cyberspace itself, not the actors using cyberspace, this same insight can be extended to AI agent design. To generalize beyond the cyberspace metaphor for which Professor Lessig's framework was originally developed, we call this approach "regulation by design" instead of regulation through "architecture."

Both companies developing AI agents and governments regulating them will have to make many design choices regarding AI agents. Many—perhaps most—of these design choices will concern specific behaviors or outcomes that we want to address. Should AI agents announce themselves as such? How frequently should they "check in" with their human principals? What sort of applications should AI agents be allowed to use?

These are all important questions. But LFAI tackles higher-order questions: How should we ensure that AI agents are regulable in general? How can we avoid creating a new class of actors unbound by law? Returning to Professor Lessig's four constraints, LFAI proposes that instead of relying solely on ex post legal sanctions, such as liability rules, we should require AI agents to be *designed to* follow some set of laws: they should be LFAIs.[226]

---

222. Now, sometimes a design requirement can be enforced with ex post damages or penalties. For example, the owner of an inaccessible building might face some sort of penalty for failing to comply with accessible design requirements. *See* 42 U.S.C. § 12183 (prohibiting as discrimination the failure to design accessible facilities); *id.* § 12188(b)(2)(A)–(B) (authorizing courts to provide money damages and fines for discrimination). Thus, ex post actions can enforce ex ante design requirements. *See* LESSIG, *supra* note 1, at 124 ("The constraints are distinct, yet they are plainly interdependent. Each can support or oppose the others."); *see also infra* Part V.B (discussing ex post strategies for enforcing LFAI design requirements).

223. *See* LESSIG, *supra* note 1, at 128.

224. *See id.* at 6 ("Code is never found; it is only ever made, and only ever made by us.").

225. *See, e.g.*, *id.* at 4 ("[T]he argument of this book is that the invisible hand of cyberspace is building an architecture that is quite the opposite of its architecture at its birth. This invisible hand, pushed by government and by commerce, is constructing an architecture that will perfect control and make highly efficient regulation possible."); *see also id.* at 15, 81, 110, 311; CHESTERMAN, *supra* note 28, at 148.

To be clear, it is not always desirable to regulate through architectural changes. *See* LESSIG, *supra* note 1, at 173–75 (expressing concerns that regulation by code will be too restrictive in the domain of copyright).

226. *See also* Robert Mahari & Alex Pentland, *Regulation by Design: A New Paradigm for Regulating AI Systems*, *in* DIGITAL SINGLE MARKET AND ARTIFICIAL INTELLIGENCE: AI ACT AND INTELLECTUAL PROPERTY IN THE DIGITAL TRANSITION 431 (Mario Franzosi, Oreste Pollicino & Gianluca Campus eds., 2024).

Thus, for whatever sets of legal constraints we wish to impose on the behavior of AI agents,[227] LFAIs will be designed to comply automatically.

B. Theoretical Motivations

1. Law Following in Principal-Agent Relationships

As discussed above,[228] AI agents can be fruitfully analyzed through principal-agent principles. Without advocating for the wholesale legal application of agency law to AI agents, reference to agency law principles can help illuminate the significance and potential of LFAI.[229]

Under hornbook agency principles, an AI agent should generally "act loyally for the principal's benefit in all matters connected with the agency relationship."[230] This generally includes a duty to obey instructions from the principal.[231]

Crucially, however, this general duty of obedience is qualified by a higher-order duty to follow the law. Agents only have a duty to obey lawful instructions.[232] Thus, "[a]n agent has no duty to comply with instructions that may subject the agent to criminal, civil, or administrative sanctions or that exceed the legal limits on the principal's right to direct action taken by the agent."[233] "[A] contract provision in which an agent promises to perform an unlawful act is unenforceable."[234] An agent cannot escape personal liability for unlawful acts ordered by their principal.[235]

The basic assumption that underlies these various doctrines is that an agent lacks any independent power to perform unlawful acts.[236] Agency law therefore developed under the assumption that agents maintain an independent obligation to follow the law and, thus, remain accountable for their violations of law. This assumption shaped agency law to prevent principals from unjustly benefiting by externalizing harms incident to the

227. *See* Question 2 of the Research Agenda, *infra* Part VI (discussing which laws LFAIs should follow).

228. *See supra* Part I.C.

229. *See supra* note 71.

230. RESTATEMENT (THIRD) OF AGENCY § 8.01 (A.L.I. 2006).

231. *See id.* § 8.09(2).

232. *See id.*

233. 3 AM. JUR. 2D *Agency* § 191 (2025); *see also* RESTATEMENT (THIRD) OF AGENCY § 8.09 cmt. c (A.L.I. 2006).

234. RESTATEMENT (THIRD) OF AGENCY § 8.09 cmt. c (A.L.I. 2006).

235. *See id.*

236. *See* 3 AM. JUR. 2D *Agency* § 53 (2025) ("Generally, the agent's authority is to perform any act the principal may lawfully perform, subject to statutory limitations and the limitations of the agreement between the principal and agent. The agent's authority is what the principal has authorized the agent to do, and the scope of the agent's authority precludes any act by an agent which the principal could not do directly or which the principal would not be authorized to do personally or does not possess the power to do." (footnotes omitted)).

agency relationship.[237] This feature of agency law helps establish a baseline to which we can compare the world of AI agents in the absence of law-following constraints; it also provides a normative justification for requiring AI agents to prioritize legal compliance over obedience to their principals.

2. Law Following in the Design of Artificial Legal Actors

AI agents will of course not be the first artificial actor that humanity has created. Two types of powerful artificial actors—corporations and governments[238]—profoundly impact our lives. When deciding how the law should respond to AI agents, it may make sense to draw lessons from the law's response to the invention of other artificial legal actors.

A key lesson for AI agents is this: for both corporations and governments, the law does not rely solely on ex post liability to steer the actor's behavior; it requires the actor to be law following by design, at least to some extent. A disposition toward compliance is built into the very "architecture" of these artificial actors. AI agents may become no less important than corporations and governments in the aggregate, not least because they will be thoroughly integrated into them. Just as the law requires these other actors to be law following by design, it should require AI agents to be LFAIs.

a. Corporations as Law Following by Design

The law requires corporations to be law following by design. One way it does this is by regulating the very legal instruments that bring corporations into existence: corporate charters are only granted for lawful purposes.[239] While an "extreme" remedy,[240] courts can order corporations to be dissolved if they repeatedly engage in illegal conduct.[241] Failure to comply with legally required corporate formalities can also be grounds for involuntarily

237. *Cf.* Paula J. Dalley, *A Theory of Agency Law*, 72 U. PITT. L. REV. 495 (2011) (explaining agency law through a cost-benefit internalization lens).

238. Since these are legal persons, they are a fortiori legal actors under our definition. *See supra* Part II.A.

239. *E.g.*, DEL. CODE. ANN. tit. 8, § 101(b) (2024) ("A corporation may be incorporated or organized under this chapter to conduct or promote any *lawful* business or purposes." (emphasis added)).

240. People v. Nat'l Rifle Ass'n of Am., Inc., 165 N.Y.S.3d 234, 249 (Sup. Ct. 2022).

241. *See, e.g.*, DEL. CODE. ANN. tit. 8, § 284(a) (2024); 12 U.S.C. § 93; Tifft v. Stevens, 987 P.2d 1, 10 (Or. Ct. App. 1999); People v. Oliver Sch., Inc., 206 A.D.2d 143, 147 (N.Y. App. Div. 1994); State v. Cortelle Corp., 38 N.Y.2d 83, 87–88 (1975). *See generally* MICHAEL CLARK, 1 CORPORATE CRIMINAL LIABILITY § 1:15 (3d ed. 2024). This is an example of an ex post procedure for enforcing a design requirement. *See supra* note 222.

dissolving a corporate entity[242] or piercing the corporate veil.[243] Thus, while corporations are, like all legal persons, generally obligated to obey the law, states do not only rely on external sanctions to persuade them to do so: they also force corporations to be law following through architectural measures, including dissolving[244] corporations that break the law or refusing to incorporate those that would.

The law also forces corporations to be law following by regulating the human agents that act on their behalf through the agents' fiduciary duties. Directors who intentionally cause a corporation to violate positive law breach their duty of good faith.[245] Not only are corporate fiduciaries required to follow the law themselves, they are required to monitor for violations of law by other corporate agents.[246] Moreover, human agents that violate certain laws can be disqualified from serving as corporate agents.[247] These sort of "structural" duties and remedies[248] are thus aimed at causing the corporation to follow the law generally and pervasively, rather than merely penalizing violations as they occur.[249] That is entirely sensible, since the state has an obvious interest in preventing the creation of new artificial entities that then go on to disregard its laws, especially since it cannot easily monitor many

242. *E.g.*, TEX. BUS. ORGS. CODE ANN. § 11.251(b) (West 2007).

243. *E.g.*, Am. Bell Inc. v. Fed'n of Tel. Workers of Pa., 736 F.2d 879, 886 (3d Cir. 1984). These are also examples of ex post procedures for enforcing a design requirement. *See supra* note 222.

244. For arguments that such remedies should be applied more frequently and aggressively than they currently are, see, for example, W. Robert Thomas, *Incapacitating Criminal Corporations*, 72 VAND. L. REV. 905 (2019); Mary Kreiner Ramirez, *The Science Fiction of Corporate Criminal Liability: Containing the Machine Through the Corporate Death Penalty*, 47 ARIZ. L. REV. 933 (2005).

245. *See, e.g.*, *In re* Walt Disney Co. Derivative Litig., 906 A.2d 27, 67 (Del. 2006) ("The good faith required of a corporate fiduciary includes not simply the duties of care and loyalty . . . but all actions required by a true faithfulness and devotion to the interests of the corporation and its shareholders. A failure to act in good faith may be shown, for instance, . . . *where the fiduciary acts with the intent to violate applicable positive law*." (emphasis added) (quoting *In re* Walt Disney Co. Derivative Litig., 907 A.2d 693, 755 (Del. Ch. 2005)).

246. *See, e.g.*, *In re* Caremark Int'l Inc. Derivative Litig., 698 A.2d 959, 970 (Del. Ch. 1996).

247. *See, e.g.*, 15 U.S.C. §§ 77t(e), 78u(d)(2) (publicly traded companies); Spencer Weber Waller, *Corporate Governance and Competition Policy*, 18 GEO. MASON L. REV. 833, 865–67 (2011); Rohit Chopra, *Reining in Repeat Offenders*, 11 REGUL. REV. 9, 18 (2022); Philip F.S. Berg, *Unfit to Serve: Permanently Barring People from Serving as Officers and Directors of Publicly Traded Companies After the Sarbanes-Oxley Act*, 56 VAND. L. REV. 1871 (2019).

248. Chopra, *supra* note 247, at 16.

249. *See, e.g.*, Mary Jo White, Chair, Sec. & Exch. Comm'n, Understanding Disqualifications, Exemptions and Waivers Under the Federal Securities Laws, Remarks at the Corporate Counsel Institute, Georgetown University in Washington D.C. (Mar. 12, 2015), https://www.sec.gov/newsroom/speeches-statements/031215-spch-cmjw [https://perma.cc/UH9S-RVPE] ("Disqualifications guard against future participation in certain capital market activities by entities or individuals whose misconduct suggests that they cannot be relied upon to conduct those activities in compliance with the law and in a manner that will protect investors and our markets." (footnote omitted)).

corporate activities. Whether a powerful and potentially difficult-to-monitor AI agent is generally disposed toward lawfulness will be similarly important. Accordingly, there is a parallel case for requiring the principals of AI agents to demonstrate that their agents will be law following.[250]

b. Governments as Law Following by Design

"Constitutionalism is the idea . . . that government can and should be legally limited in its powers, and that its authority or legitimacy depends on its observing these limitations."[251] Although we sometimes rely on ex post liability to deter harmful behavior by government actors,[252] the design of the government—through the U.S. Constitution,[253] statutory provisions, and longstanding practice—is the primary safeguard against lawless government action.

Examples abound. The general American constitutional design of separated powers, supported by interbranch checks and balances, plays an important role in preventing the government from exercising arbitrary power, thereby confining the government to its constitutionally delimited role.[254] This system of multiple independent veto points yields concrete protections for personal liberty, such as making it difficult for the government to lawlessly imprison people.[255]

Governments, like corporations, act only through their human agents.[256] As in the corporate case, governmental design forces the branches of government to follow the law in part by imposing law-following duties on the agents through whom it acts. The Constitution imposes a duty on the president to "take Care that the Laws be faithfully executed."[257] Similar to

250. *Cf.* CHOPRA & WHITE, *supra* note 28, at 167–68 ("In principle, artificial agents could also be restrained by purely technical means, by being disabled, or banned from engaging in economically rewarding work for stipulated periods. . . . Deregistration of an agent or confiscation of its assets might also be used as a sanction, just as winding-up is used to end the life of companies in certain situations, or confiscation is used concerning the proceeds of crime.").

251. Wil Waluchow & Dimitrios Kyritsis, *Constitutionalism*, STAN. ENCYC. PHIL. (May 18, 2023), https://plato.stanford.edu/entries/constitutionalism/index.html [https://perma.cc/Z9VW-GUGQ].

252. *See, e.g.*, 42 U.S.C. § 1983.

253. *Cf.* LESSIG, *supra* note 1, at 6 (describing the Constitution as a design constraint on government).

254. *See, e.g.*, MONTESQUIEU, THE SPIRIT OF THE LAWS, bk. XI ch. VI (1748); THE FEDERALIST NOS. 47, 48 (James Madison).

255. Brief of Akhil Reed Amar & Vikram David Amar as Amici Curiae in Support of Neither Party at 26, Trump v. Anderson, 144 S. Ct. 662 (2024) (No. 23-719).

256. *E.g.*, Mortimer N.S. Sellers, *International Legal Personality*, 11 IUS GENTIUM 67, 70 ("States . . . act, if they act at all, through and upon real persons.").

257. U.S. CONST. art. II, § 3; *see also* Mary M. Cheh, *When Congress Commands a Thing to Be Done: An Essay on* Marbury v. Madison*, Executive Inaction, and the Duty of the Courts to Enforce the Law*, 72 GEO. WASH. L. REV. 253, 275 (2003) (explaining that the Take Care Clause "is a *duty* imposed on the president, not a source of power per se"). *See generally* Jack

the corporate context discussed above, soldiers have a duty to disobey some unlawful orders, even from the commander in chief.[258] Civil servants also have a right to refuse to follow unlawful orders, though the exact nature and extent of this right is unclear.[259]

We saw above that, in the corporate case, the law uses disqualification of law-breaking agents to ensure that corporations are law following.[260] The law also uses disqualification to ensure that the government acts only through law-following agents, ranging from the highest levels of government to lower-level bureaucrats and employees. The Constitution empowers Congress to remove and disqualify officers of the United States for "high Crimes and Misdemeanors" through the impeachment process.[261] Each house of Congress may expel its own members for "disorderly Behaviour."[262] Historically, Congress has exercised this power in cases "involv[ing] either disloyalty to the United States Government, or the violation of a criminal law involving the abuse of one's official position, such as bribery."[263] Although there is no blanket rule disqualifying persons with criminal records from federal government jobs,[264] numerous laws disqualify convicted individuals in more specific circumstances.[265] Convicted felons are also generally ineligible to be employed by the Federal Bureau of Investigation[266] or armed forces[267] and usually cannot obtain a security clearance.[268]

These design choices encode a commonsense judgment that those who cannot be trusted to follow the law should not be entrusted to wield the extraordinary power that accompanies certain government jobs, especially positions associated with law enforcement, the military, and the intelligence community. If AI agents come to wield similar power and influence, the case for designing them to obey the law is equally compelling.

Goldsmith & John F. Manning, *The Protean Take Care Clause*, 164 U. PENN. L. REV. 1835 (2016).

258. *See supra* note 115.

259. *See, e.g.*, Alex Hemmer, Note, *Civil Servant Suits*, 124 YALE L.J. 758, 785–90 (2014); Robert G. Vaughn, *Public Employees and the Right to Disobey*, 29 HASTINGS L.J. 261 (1977).

260. *See supra* note 247 and accompanying text.

261. U.S. CONST. art. 1, §§ 2–3; *id.* art. II, § 4; *see also id.* art. III, § 1 (limiting judges' tenure in office to "good Behaviour").

262. *Id.* art. I, § 5.

263. JACK MASKELL, CONG. RSCH. SERV., RL31382, EXPULSION, CENSURE, REPRIMAND, AND FINE: LEGISLATIVE DISCIPLINE IN THE HOUSE OF REPRESENTATIVES 3 (2005).

264. *See, e.g.*, DEP'T OF JUST., FEDERAL STATUTES IMPOSING COLLATERAL CONSEQUENCES UPON CONVICTION 3, https://www.justice.gov/sites/default/files/pardon/legacy/2006/11/13/collateral_consequences.pdf [https://perma.cc/9EYS-ACC3].

265. *Id.* at 2–3.

266. *See Eligibility and Hiring: What It Takes to Join the FBI*, FED. BUREAU OF INVESTIGATION JOBS, https://fbijobs.gov/eligibility [https://perma.cc/886A-TAYW] (last visited Aug. 1, 2025).

267. *See* 10 U.S.C. § 504(a).

268. *See, e.g.*, 32 C.F.R. § 147.12 (1998).

3. The Holmesian Bad Man and the Internal Point of View

Our distinction between AI henchmen and LFAIs mirrors a distinction in jurisprudence about possible attitudes toward legal obligations.[269] An AI henchman treats legal obligations much as the "bad man" does in Justice Oliver Wendell Holmes Jr.'s classic *The Path of the Law*:

> If you want to know the law and nothing else you must look at it as a bad man, who cares only for the material consequences which such knowledge enables him to predict, not as a good one, who finds his reasons for conduct, whether inside the law or outside of it, in the vaguer sanctions of conscience.[270]

That is, under some interpretations,[271] Justice Holmes' bad man treats the law merely as a set of incentives within which he pursues his own self-interest.[272] Like the bad man, the primary reason an AI henchman would have to follow the law is simply that, if it is caught breaking the law, its principal might face negative consequences that outweigh the benefits gained from the lawbreaking.[273] Like the bad man,[274] therefore, if the AI henchman predicts that the expected costs of violating the law are greater than the expected benefits, it will obey. Otherwise, it will not.

Fortunately, the bad man is not the only possible model for AI agents' attitudes toward the law. One alternative to the bad man view of the law is Professor H.L.A. Hart's "internal point of view."[275] "The internal point of view is the practical attitude of rule acceptance—it does not imply that people

---

269. Thank you to Professor Peter Salib for first making this point to one of us in conversation. Two other sources make arguments similar to those in this part: Nerantzi & Sartor, *supra* note 36, at 698–700, and Lemley & Casey, *supra* note 28, at 1345–51.

270. Oliver Wendell Holmes Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 459 (1897).

271. *See* Adrian Vermeule, *"Above the Law"*, THE NEW DIGEST (July 3, 2024), https://thenewdigest.substack.com/p/above-the-law [https://perma.cc/R25G-BEJX] ("*[I]n some interpretations [of Holmes]* anyway, only the threat of coercive legal process, civil or criminal, will induce the bad man to obey the law" (emphasis added)). For another interpretation, see David Luban, *The Bad Man and the Good Lawyer*, *in* THE PATH OF THE LAW AND ITS INFLUENCE 33 (Steven J. Burton ed., 2000), who distinguishes the Holmesian bad man from the "*very* bad man, who violates a contract and refuses to pay the fine and obstructs the effort to collect the fine." *Id.* at 40. Under this reading, AI henchmen are closer to the very bad man, not Justice Holmes's bad man.

272. *E.g.*, Robert W. Gordon, *The Path of the Lawyer*, 110 HARV. L. REV. 1013, 1014 (1997).

273. *Cf.* Claire Finkelstein, *Hobbes and the Internal Point of View*, 75 FORDHAM L. REV. 1211, 1213 (2006) ("[The bad man] has no sense of legal duty and would think nothing of violating the law if he could do so with impunity. The only restraint on illegality is the possibility of detection, which he would constantly weigh against the potential for gain.").

274. *Cf.* Albert W. Alschuler, *The Descending Trail: Holmes'* Path of the Law *One Hundred Years Later*, 49 FLA. L. REV. 353, 375 n.84 (1997) ("Once people internalize the 'bad man' perspective, the assumption that easily evaded law is not law becomes routine. Because that is what Holmes' definition says, it encourages the view that taking advantage of loopholes is unproblematic and that nearly everyone will do so.").

275. H.L.A. HART, THE CONCEPT OF LAW 89 (2d ed. 1994).

who accept the rules accept their moral legitimacy, only that they are disposed to guide and evaluate conduct in accordance with the rules."[276] Whether AIs can have the capacity to truly adopt the internal point of view is, of course, contested.[277] But regardless of their mental state (if any), AI agents can be designed to act like someone who thinks that "the law is not simply sanction-threatening, -directing, or -predicting, but rather obligation-imposing"[278] and is thus disposed to "act[] according to the dictates of the [law]."[279] An AI agent can be designed to be more rigorously law-following than the bad man.[280]

Real life is of course filled with people who are "bad" or highly imperfect. But bad AI agents are not similarly inevitable. AI agents are human-designed artifacts. It is open to us to design their behavioral dispositions to suit our policy goals, and to refuse to deploy agents that do not meet those goals.

C. Concrete Benefits

1. Law-Following AI Prevents Abuses of Government Power

As we have discussed,[281] the law makes the government follow the law (and thus prevents abuses of government power) in part by compelling government *agents* to follow the law. If the government comes to rely heavily on AI agents for cognitive labor, then the law should also require those agents to follow the law.

Depending on their assigned "roles," government AI agents could wield significant power. They may have authority to initiate legal processes against individuals (including subpoenas, warrants, indictments, and civil actions), access sensitive governmental information (including tax records and intelligence), hack into protected computer systems, determine eligibility for government benefits, operate remote-controlled vehicles like military drones,[282] and even issue commands to human soldiers or law enforcement officials.

These powers present significant opportunities for abuse, which is why preventing lawless government action was a motivation for the American Revolution,[283] a primary goal of the Constitution, and a foundational American political value. We must therefore carefully examine whether existing safeguards designed to constrain human government agents would

276. Scott J. Shapiro, *What Is the Internal Point of View?*, 75 FORDHAM L. REV. 1157, 1157 (2006).
277. *See supra* Part II.B.
278. Shapiro, *supra* note 276, at 1157.
279. *Id.* at 1162.
280. For empirical support that normative acceptance of the law's content is an important driver of compliance, see TYLER, *supra* note 76.
281. *See supra* Part III.B.2.b.
282. *See supra* note 56.
283. *See, e.g.*, THE DECLARATION OF INDEPENDENCE pmbl. (U.S. 1776).

effectively limit AI agents in the absence of the law-following design constraints. While our analysis here is necessarily incomplete, we think it provides some reason for doubting the adequacy of existing safeguards in the world of AI agents.

When a human government agent, acting in their official capacity, violates an individual's rights, they can face a variety of ex post consequences. If the violation is criminal, they could face severe penalties.[284] This "threat of criminal sanction for subordinates [i]s a very powerful check on executive branch officials."[285] The threat of civil suits seeking damages, such as through a § 1983[286] or *Bivens* action,[287] might also deter them, though various immunities and indemnities will often protect them,[288] especially if they are a federal officer.[289]

These checks will not exist in the case of AI henchmen. In the absence of law-following constraints, an AI henchman's primary reason to obey the law will be its desire to keep its *principal* out of trouble.[290] The henchman will

284. "[T]he two primary statutes that criminalize the actions of governmental officials who abuse their authority to deprive their fellow citizens of their constitutional rights" are 18 U.S.C. §§ 241–42. Samantha Trepel, *Prosecuting Color-of-Law Civil Rights Violations: A Legal Overview*, 70 DOJ J. FED. L. & PRAC. 21, 21 (2022). Section 241 criminalizes "conspir[acy] to injure, oppress, threaten, or intimidate any person in [the United States] in the free exercise or enjoyment of any right or privilege secured to him by the Constitution or laws of the United States, or because of his having so exercised the same." 18 U.S.C. § 241. Section 242 makes it a crime to, inter alia, "willfully subject[] any person in [the United States] to the deprivation of any rights, privileges, or immunities secured or protected by the Constitution or laws of the United States." *Id.* § 242. Penalties for both statutes include, at their most extreme, death. *See id.* §§ 241, 242.

285. Jack Goldsmith, *The Relative Insignificance of the Immunity Holding in* Trump v. United States *(and What Is Really Important in the Decision)*, LAWFARE (Sep. 23, 2024, 12:52), https://www.lawfaremedia.org/article/the-relative-insignificance-of-the-immunity-holding-in-trump-v.-united-states-(and-what-is-really-important-in-the-decision) [https://perma.cc/2DW6-YEC6].

286. 42 U.S.C. § 1983 (authorizing civil suits against employees of state governments).

287. Bivens v. Six Unknown Named Agents, 403 U.S. 388 (1971) (authorizing civil suits against employees of the federal government). Of course, the Supreme Court has since dramatically limited the availability of *Bivens* actions. *See, e.g.*, Henry Rose, *The Demise of the* Bivens *Remedy Is Rendering Enforcement of Federal Constitutional Rights Inequitable but Congress Can Fix It*, 42 N. ILL. U. L. REV. 229, 232–37 (2022).

288. *See, e.g.*, Harlow v. Fitzgerald, 457 U.S. 800 (1982) (establishing the modern test for qualified immunity); Joanna C. Schwartz, *Police Indemnification*, 89 N.Y.U. L. REV. 885, 890–91 (2014) (documenting near-universal indemnification of law enforcement officers).

289. *See, e.g.*, 28 U.S.C. § 2679(b) (codifying part of the Federal Employees Liability Reform and Tort Compensation Act of 1988, Pub. L. No. 100-694, 102 Stat. 4563, also known as the Westfall Act; substituting the United States as the defendant in civil suits against federal employees for actions within the scope of their office, except for violations of the Constitution and explicitly granted statutory causes of action). "Scholars have long assumed that the Westfall Act immunity broadly immunizes federal employees for wrongful acts committed within the scope of their employment." James E. Pfander & Rex N. Alley, *Federal Tort Liability After* Egbert v. Boule*: The Case for Restoring the Officer Suit at Common Law*, 138 HARV. L. REV. 985, 990 n.25 (2025); *see also* United States v. Smith, 499 U.S. 160, 166 (1991).

290. *See supra* Part I.C. A court may of course try issuing an injunction that requires a principal to prevent their AI agent from taking certain illegal actions. However, unless the AI

thus lack one of the most powerful constraints on lawless behavior in humans: fear of personal ex post liability.

Most of us would rightfully be terrified of a government staffed by agents whose only concern was whether their bosses would suffer negative consequences as a result of their actions—a government staffed by Holmesian bad men loyal only to their principals.[291] A basic premise of American constitutionalism,[292] and rule-of-law principles more generally,[293] is that government officials act legitimately only when they act pursuant to powers the people granted to them through law and obey the constraints attached to those powers. Treating law as a mere incentive system is repugnant to the proper role of government agents:[294] being a "servant" of the people[295] "faithfully discharg[ing] the duties of [one's] office."[296]

This is not just a matter of high-minded political and constitutional theory. An elected head of state aspiring to become a dictator would need the cooperation of the sources of hard power in society—military, police, other security forces, and government bureaucracy—to seize power. At present, however, these organs of government are staffed by individuals, who may choose not to go along with the aspiring dictator's plot.[297] Furthermore, in an economy dependent on diffuse economic activity, resistance by individual workers could reduce the economic upsides of a coup.[298] This reliance on a diverse and imperfectly loyal human workforce, both within and outside of

---

agent is designed to follow such injunctions (i.e., unless they are LFAIs), this does not change the fundamental calculus: if the agent can find a way to benefit its principal without putting them at risk of contempt, it will do so. *Cf. supra* note 80 and accompanying text (discussing how AI henchmen might hide lawbreaking behavior from their principals to preserve the principals' plausible deniability).

291. *Cf.* Vermeule, *supra* note 271 (saying that a vision of government wherein "public magistrate[s only] obey[] the law . . . in the shadow of coercion" is "a public-law version of Holmes' 'bad man' theory of the law, according to which, in some interpretations anyway, only the threat of coercive legal process, civil or criminal, will induce the bad man to obey the law" (citing Holmes, *supra* note 270)).

292. *See, e.g.*, THE DECLARATION OF INDEPENDENCE para. 2 (U.S. 1776) ("Governments are instituted among Men, deriving their just powers from the consent of the governed."); THE FEDERALIST NO. 78 (Alexander Hamilton) ("There is no position which depends on clearer principles, than that every act of a delegated authority, contrary to the tenor of the commission under which it is exercised, is void. No legislative act, therefore, contrary to the Constitution, can be valid. To deny this, would be to affirm . . . that men acting by virtue of powers, may do not only what their powers do not authorize, but what they forbid.").

293. *See generally* Jeremy Waldron, *The Rule of Law*, STAN. ENCYC. PHIL. (2023), https://plato.stanford.edu/entries/rule-of-law [https://perma.cc/GZ5X-QKTB].

294. As discussed in Part III.B.1 above, it is also inconsistent with general fiduciary principles.

295. *See generally* THE FEDERALIST NO. 78 (Alexander Hamilton).

296. 5 U.S.C. § 3331 (oath of office for government officials other than the president); *see also* U.S. CONST. art. II, § 1, cl. 8 (President swears an oath to "faithfully execute the Office of President of the United States"); *supra* Parts III.B.1, III.B.2 (discussing law following as a matter of fiduciary principle).

297. *See generally* Moritz von Knebel, Tearing at the Seams?: Automation and the Decline of Democracy 3–8 (unpublished manuscript) (on file with the *Fordham Law Review*).

298. *See id.*

government, is a significant safeguard against tyranny.[299] However, replacement of human workers with loyal AI henchmen would seriously weaken this safeguard, possibly easing the aspiring tyrant's path to power.[300]

Nor is the importance of LFAIs limited to AI agents acting directly at the request of high-level officials. It extends to the vast array of lower-level state and federal officials who wield enormous power over ordinary citizens, including particularly powerless ones. Take prisons, which "can often seem like lawless spaces, sites of astonishing brutality where legal rules are irrelevant."[301] Prison law arguably constrains abuse by officials far less than it should. Nevertheless, "prisons are intensely legal institutions," and "people inside prisons have repeatedly emphasized that legal rules have significant, concrete effects on their lives."[302] Even imperfect enforcement of the legal constraints on prison officials can have demonstrable effects.[303] However bad the existing situation may be, diluting or gutting the efficacy of these constraints threatens to make the situation dramatically worse.

The substitution of AI agents for (certain) prison officials could have precisely this effect. Here is just one example. The Eighth Amendment forbids prison officials from withholding medical treatment from prisoners in a manner that is deliberately indifferent to their serious medical needs.[304] Suppose that a state prisoner needs to take a dose of medicine each day for a month or their eyesight will be permanently damaged. The prisoner says something disrespectful to a guard. The warden, hoping to make an example of the prisoner, fabricates a note from the prison physician directing the prison pharmacist to withhold further doses of the medicine. The prisoner is subsequently denied the medicine. They try to reach their lawyer to get a temporary restraining order, but the lawyer cannot return their call until the next day. As a result, the prisoner's eyesight is permanently damaged.

---

299. *See id.*

300. *See, e.g.*, *id.* at 17–18; Lee Drutman & Yascha Monk, *Will Robots Kill Democracy*, NAT'L INTEREST, July/Aug. 2016, at 22; Matthew M. Young, Johannes Himmelreich, Justin B Bullock & Kyoung-Cheol Kim, *Artificial Intelligence and Administrative Evil*, 4 PERSP. ON PUB. MGMT. & GOVERNANCE 244 (2021); Christoph K. Winter, *The Challenges of Artificial Judicial Decision-Making for Liberal Democracy*, *in* JUDICIAL DECISION-MAKING: INTEGRATING EMPIRICAL AND THEORETICAL PERSPECTIVES 179, 195–98 (Piotr Bystranowski, Bartosz Janik & Maciej Próchnicki eds., 2022); Dan Hendrycks, Mantas Mazeika & Thomas Woodside, An Overview of Catastrophic AI Risks 10 (Oct. 23, 2023) (unpublished manuscript), https://arxiv.org/pdf/2306.12001 [https://perma.cc/8K3J-KJTU]; *cf.* Bryan Caplan, *The Totalitarian Threat*, *in* GLOBAL CATASTROPHIC RISKS 449, 454–55, 457–58 (Milan M. Ćirković & Nick Bostrom eds., 2008) (discussing new technologies as a risk factor for global totalitarianism).

301. Emma Kaufman & Justin Driver, *The Incoherence of Prison Law*, 135 HARV. L. REV. 515, 520 (2021).

302. *Id.* at 521.

303. *Id.*

304. *See* Estelle v. Gamble, 429 U.S. 97, 104–05 (1976).

Let us assume that the state has strong state-level sovereign immunity under its own laws, meaning that the prisoner cannot sue the state directly.[305] Under the status quo, the prisoner can still sue the warden for damages under 42 U.S.C. § 1983 for violating their clearly established constitutional right.[306] Given the widespread prevalence of official indemnification agreements at the state level,[307] the state will likely indemnify the warden, even though the state itself cannot be sued for damages under § 1983[308] or its own laws. The prisoner is therefore likely to receive monetary damages.

But now replace the human warden with an AI agent charged with administering the prison by issuing orders directly to prison personnel through some digital interface. If this "AI warden" did the same thing, the prisoner would not have direct redress against it, since it is not a "person" under § 1983[309] (or, indeed, any law). Nor will the prisoner have indirect recourse against the state by way of an indemnification agreement because there is no underlying tort liability for the state to indemnify. Nor will the prisoner have redress against the medical personnel, since the AI warden deceived them into withholding treatment.[310] And, as we have already assumed, the state itself has sovereign immunity. Thus, the prisoner will find themself without any avenue of redress for the wrong they have suffered, and the introduction of an artificial agent in the place of a human official made all the difference.

What is the right response to these problems? Many responses may be called for, but one of them is to ensure that only law-following AI agents can serve in such a role. As previously discussed, the law disqualifies certain lawbreakers from many government jobs. Similarly, we believe, the law should disqualify AI agents that are not demonstrably rigorously law following from certain government roles. We discuss how this disqualification might be enforced, more concretely, in Part V.

There is, however, another possible response to these challenges: perhaps we should "just say no" and prohibit governments from using AI agents at all or at least severely curtail their use.[311] We do not take a strong position

305. We will assume for the purposes of this hypothetical that the state is one like Alabama, with very strong sovereign immunity. *See* Hutchinson v. Bd. of Trs. of Univ. of Ala., 288 Ala. 20, 24 (1971) (explaining that Alabama state sovereign immunity is "almost invincible").

306. *Estelle* clearly established that "intentionally interfering with . . . treatment once prescribed" violates the Eighth Amendment. 429 U.S. at 104–05. This "existing precedent [has] placed the . . . constitutional question beyond debate." Ashcroft v. al-Kidd, 563 U.S. 731, 741 (2011). Thus, the warden likely would not enjoy qualified immunity. *See, e.g.*, Williams v. Treen, 671 F.2d 892, 901 (5th Cir. 1982).

307. *See generally* Schwartz, *supra* note 288.

308. *See* Edelman v. Jordan, 415 U.S. 651, 667–69 (1974).

309. 42 U.S.C. § 1983 only creates liability for "person[s]." 42 U.S.C. § 1983.

310. Inadvertent or accidental failures to administer necessary medical care do not ordinarily give rise to a § 1983 claim. *See Estelle*, 429 U.S. at 105–06.

311. *Cf.* Caplan, *supra* note 300, at 458 ("[T]he safest approach [to avoid totalitarianism] is freedom for individuals combined with heavy scrutiny for government. In the hands of individuals, new technology helps people pursue their diverse ends more effectively. In the

on when this would be the correct approach, all things considered. At a minimum, however, we note a few reasons for skepticism of such a restrictive approach.

The first is banal: if AI agents can perform computer-based tasks well, then their adoption by the government could also deliver considerable benefits to citizens.[312] Reducing the efficiency of government administration for the sake of preventing tyranny and abuse may be worthwhile in some cases and is indeed the logic of the individual rights protections of the Constitution.[313] But tailoring a safeguard to allow for efficient government administration, is, all else being equal, preferable to a blunter, more restrictive safeguard. LFAI may offer such a tailored safeguard.

The second reason for skepticism is that adoption of AI agents by governments may become more important as AI technology advances. Some of the most promising AI safety proposals involve using trusted AI systems to monitor untrusted ones.[314] The central reason is this: as AI systems become more capable, unassisted humans will not be able to reliably evaluate whether the AIs' actions are desirable.[315] Assistance from trusted AI systems could thus be the primary way to scale humans' ability to oversee untrusted AI systems. If the government is to oversee the behavior of new and untrusted private-sector AI systems, it might be necessary to do so using AI agents.

Even if the government does not need to rely on AI agents to administer AI safety regulation (for example, because such AI overseers are employed by private companies, not the government), the government will likely need to employ AI agents to help it keep up with competitive pressures. Should the federal government hesitate to adopt AI agents to increase its efficiency, foreign competitors might show no such qualms. As a result, the federal government might then feel little choice but to likewise adopt AI agents.

---

hands of government, however, new technology risks putting us on the slippery slope to totalitarianism.").

312. *Cf.* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine Learning Era*, 105 GEO. L.J. 1147, 1148 (2017) (urging "measured optimism about the potential benefits [algorithmic automation] technology can offer society by making government smarter and its decisions more efficient and just").

313. *See supra* notes 254–55 and accompanying text; SEC v. Jarkesy, 144 S. Ct. 2117, 2149–50 (2024) (Gorsuch, J., concurring).

314. *See, e.g.*, Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan & Fabien Roger, *AI Control: Improving Safety Despite Intentional Subversion*, *in* PROC. 41ST INT'L CONF. ON MACH. LEARNING (2024), https://proceedings.mlr.press/v235/greenblatt24a.html (on file with the *Fordham Law Review*); Jan Leike, *Why I'm Excited About AI-Assisted Feedback*, MUSINGS ON THE ALIGNMENT PROBLEM (Mar. 29 2022), https://aligned.substack.com/p/ai-assisted-human-feedback [https://perma.cc/5ZS2-RUUY]; *see also* Etzioni & Etzioni, *supra* note 85, at 139.

315. *See, e.g.*, Greenblatt et al., *supra* note 314, at 1; Richard Ngo, Lawrence Chan & Sören Mindermann, *The Alignment Problem from a Deep Learning Perspective*, *in* PROC. 12TH INT'L CONF. ON LEARNING REPRESENTATIONS 3–5 (2024), https://openreview.net/forum?id=fh8EYKFKns (on file with the *Fordham Law Review*).

In the face of these competing demands, LFAI offers a plausible path to enable the adoption of AI agents in governmental domains with a high potential for abuse (e.g., the military, intelligence, law enforcement, prison administration) while safeguarding life, liberty, and the rule of law. LFAI can also transform the binary question of whether to adopt AI agents into the more multidimensional question of which laws should constrain them.[316] This should allow for more nuanced policymaking, grounded in the existing legal duties of government agents.

2. Law-Following AI Enables Scalable Enforcement of Public Law

AI agents could cause a wide variety of harms. The state promulgates and enforces public law prohibitions—both civil and criminal—to prevent and remedy many of these harms. If the state cannot safely assume that AI agents will reliably follow these prohibitions, the state might need to increase the resources dedicated to law enforcement.

LFAI offers a way out of this bind. Insofar as AI agents are reliably law following, the state can trust that significantly less law enforcement is needed.[317] This dynamic would also have broader beneficial implications for the structure and functioning of government. "If men were angels, no government would be necessary."[318] LFAIs would not be angels,[319] but they would be a bit more angelic than many humans. Thus, as a corollary of Publius's insight, we may need less government to oversee LFAIs' behavior than we would need for a human population of equivalent size. State resources that would otherwise be spent on investigating and enforcing the laws against AI agents could instead be directed to other problems or refunded to the citizenry.

LFAI would also curtail some of the undesirable side effects and opportunities for abuse inherent in law enforcement. Law enforcement efforts often involve some intrusion into the private affairs and personal freedoms of citizens.[320] If the government could be more confident that AI agents under private control were behaving lawfully, it would have less cause

316. *See* Etzioni & Etzioni, *supra* note 85, at 142–44.

317. Many of the considerations raised here also imply that LFAI could lower costs of enforcing private contracts and treaties. *See* Cullen O'Keefe, *Law-Following AI 3: Lawless AI Agents Undermine Stabilizing Agreements*, ALIGNMENT F. (Apr. 28, 2022), https://www.alignmentforum.org/s/ZytYxd523oTnBNnRT/p/DfcXaGH7XGYjW22C2 [https://perma.cc/F9K2-PYVV].

318. THE FEDERALIST NO. 51 (James Madison).

319. That is, LFAIs would not be morally perfect beings. This is for at least two reasons. First, many immoral actions are legal, often rightfully so. Second, sometimes immoral actions can be legally compulsory—everyone can surely think of a tax-funded governmental program which they believe to be immoral to support. On the reasons to prefer LFAI over alignment to extralegal moral values, see *infra* Part IV.B.

320. *See* Etzioni & Etzioni, *supra* note 85, at 142–44.

to surveil or investigate their behavior, thereby imposing fewer[321] burdens on their principals' privacy. Reducing the occasion for investigations and searches would also create fewer opportunities for abuse of private information.[322] In this way, ensuring reliably law-following AI might significantly mitigate the frequency and severity of law enforcement's intrusions on citizens' privacy and liberty.

IV. LAW-FOLLOWING AI AS AI ALIGNMENT

The field of AI alignment aims to ensure that powerful, general-purpose AI agents behave in accordance with some set of normative constraints.[323] AI systems that do not behave in accordance with such constraints are said to be "misaligned" or "unaligned." Since the law is a set of normative constraints, the field of AI alignment is highly relevant to LFAI.[324]

The most basic set of normative constraints to which an AI could be aligned is the "informally specified"[325] intent of its principal.[326] This is called "intent-alignment."[327] Since individuals' intentions are a mix of morally good and bad to varying degrees, some alignment work also aims to ensure that AI systems behave in accordance with moral constraints,

321. Since the law is open-textured, and perfect enforcement of law is not desirable, even LFAIs may occasionally break the law. *See* Question 7 of the Research Agenda, *infra* Part VI. Thus, there will likely remain some need for investigations of LFAIs.

322. *Cf.* Etzioni & Etzioni, *supra* note 85, at 142–44 (discussing the detrimental effects aggressive AI-enabled investigations could have on civil liberties); J.P. de Mello Barreto, *Regulatory Models for Monitoring AI and Protecting Privacy*, INST. FOR L. & A.I. (Mar. 25, 2025), https://law-ai.org/balancing-safety-and-privacy-regulatory-models-for-ai-misuse/ [https://perma.cc/Y95W-VQSV] (discussing the Fourth Amendment implications of AI-enabled surveillance).

323. *See, e.g.*, Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACH. 411, 413 (2020); Ngo, Chan & Mindermann, *supra* note 315, at 1.

324. For other work exploring the relationship between AI alignment and law, see John J. Nay, *Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans*, 20 NW. J. TECH. & INTELL. PROP. 309 (2023); Nicholas A. Caputo, *Alignment as Jurisprudence*, YALE J.L. & TECH. (forthcoming), https://ssrn.com/abstract=4800894 (on file with the *Fordham Law Review*).

325. Ngo, Chan & Mindermann, *supra* note 315, at 1. This condition is important because it is very hard for humans to formally specify all of what we want: a problem well known to lawyers. *See* Hadfield-Menell & Hadfield, *supra* note 71.

326. *See, e.g.*, Ajeya Cotra, *"Aligned" Shouldn't Be a Synonym for "Good"*, PLANNED OBSOLESCENCE (Mar. 26, 2023), https://www.planned-obsolescence.org/aligned-vs-good/ [https://perma.cc/L34Z-BGRZ].

327. *See, e.g.*, Paul Christiano, *Clarifying "AI Alignment"*, AI ALIGNMENT (Apr. 8, 2018), https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6 [https://perma.cc/N3YX-MWKM].

regardless of the intentions of the principal.[328] This is called "value-alignment."[329]

AI alignment work is valuable because, as shown by theoretical arguments[330] and empirical observations,[331] it is difficult to design AI systems that reliably obey any particular set of constraints provided by humans.[332] In other words, nobody knows how to ensure that AI systems are either intent-aligned or value-aligned,[333] especially for smarter-than-human systems.[334] This is the "Alignment Problem."[335] The Alignment Problem is especially worrying for AI systems that are agentic and goal-directed,[336] as such systems may try to evade human oversight and controls that could frustrate pursuit of those goals, such as by deceiving their developers,[337]

328. *See, e.g.*, Christoph Winter, Nick Hollman & David Manheim, *Value Alignment for Advanced Artificial Judicial Intelligence*, 60 AM. PHIL. Q. 187 (2022); Gabriel, *supra* note 323, at 422–24.

329. *See, e.g.*, Iason Gabriel & Vafa Ghazavi, *The Challenge of Value Alignment: From Fairer Algorithms to AI Safety*, *in* THE OXFORD HANDBOOK OF DIGITAL ETHICS 336 (Carissa Véliz ed., 2021); Winter, Hollman & Manheim, *supra* note 328.

330. *See, e.g.*, Ngo, Chan & Mindermann, *supra* note 315; Michael K. Cohen, Marcus Hutter & Michael A. Osborne, *Advanced Artificial Agents Intervene in the Provision of Reward*, 43 A.I. MAG. 282 (2022); Yohan J. John, Leigh Caldwell, Dakota E. McCoy & Oliver Braganza, *Dead Rats, Dopamine, Performance Metrics, and Peacock Tails: Proxy Failure Is an Inherent Risk in Goal-Oriented Systems*, BEHAV. & BRAIN SCI., June 2023, at 1; Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel & Stuart Russell, *The Off-Switch Game*, *in* PROC. 26TH INTL. JOINT CONF. ON A.I. 220 (2017), https://www.ijcai.org/proceedings/2017/0032.pdf [https://perma.cc/6779-WVR2].

331. *See generally, e.g.*, Leonard Dung, *Current Cases of AI Misalignment and Their Implications for Future Risks*, SYNTHESE, Oct. 2023, at 1, 8–10; Ryan Greenblatt et al., Alignment Faking in Large Language Models (Dec. 20, 2024) (unpublished manuscript), https://arxiv.org/pdf/2412.14093 [https://perma.cc/4L6S-C62C]; Jack Clark & Dario Amodei, *Faulty Reward Functions in the Wild*, OPENAI (Dec. 21, 2016), http://openai.com/index/faulty-reward-functions/ [https://perma.cc/YYQ3-D7E6]; *Sycophancy in GPT-4o: What Happened and What We're Doing About It*, OPENAI (Apr. 29, 2025), https://openai.com/index/sycophancy-in-gpt-4o/ [https://perma.cc/D5J5-JA9V].

332. *See, e.g.*, STUART RUSSELL, HUMAN COMPATIBLE (2019); DAN HENDRYCKS, INTRODUCTION TO AI SAFETY, ETHICS, AND SOCIETY § 1.5 (2024).

333. *See, e.g.*, Dan Hendrycks, Nicholas Carlini, John Schulman & Jacob Steinhardt, Unsolved Problems in ML Safety 2 (June 16, 2022) (unpublished manuscript), https://arxiv.org/pdf/2109.13916 [https://perma.cc/H539-E4M6].

334. *See, e.g.*, Jan Leike, *What is the Alignment Problem?*, MUSINGS ON THE ALIGNMENT PROBLEM (Mar. 29, 2022), https://aligned.substack.com/p/what-is-alignment [https://perma.cc/C2MF-RAJT]; Jan Leike & Ilya Sutskever, *Introducing Superalignment*, OPENAI (July 5, 2023), https://openai.com/index/introducing-superalignment/ [https://perma.cc/ZKF9-FBR9].

335. *E.g.*, Ngo, Chan & Mindermann, *supra* note 315, at 1.

336. *See, e.g.*, NICK BOSTROM, SUPERINTELLIGENCE 123–93 (2014); Chan et al., *supra* note 57, at 657–69; Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott & Tom Everitt, *Discovering Agents*, A.I., June 2023, at 1, 1; Toby Shevlane et al., Model Evaluations for Extreme Risks 12 (Sep. 22, 2023) (unpublished manuscript), https://arxiv.org/pdf/2305.15324 [https://perma.cc/VSP6-WMSR].

337. *See, e.g.*, HENDRYCKS, *supra* note 332, §§ 3.4.1–.4.3; BOSTROM, *supra* note 336, at 141–45; Ngo, Chan & Mindermann, *supra* note 315; Thilo Hagendorff, *Deception Abilities Emerged in Large Language Models*, PROC. NAT'L ACAD. SCI., June 2024, at 1; Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse & Scott Garrabrant, Risks from

accumulating power and resources[338] (including by making themselves smarter),[339] and ultimately resisting efforts to correct their behavior or halt further actions.[340]

There is a sizable literature arguing that these dynamics imply that misaligned AI agents pose a nontrivial risk to the continued survival of humanity.[341] The case for LFAI, however, in no way depends on the correctness of these concerns: the specter of widespread lawless AI action should be sufficient on its own to motivate LFAI. Nevertheless, the alignment literature produces several valuable insights for the pursuit of LFAI.

A. AI Agents Will Not Follow the Law by Default

The alignment literature suggests that there is a significant risk that AI agents will not be law following by default. This is a straightforward implication of the Alignment Problem. To see how, imagine a morally upright principal who intends for his AI agent to rigorously follow the law. If the AI agent was intent-aligned, the agent would therefore follow the law. But the fact that intent-alignment is an unsolved problem implies that there is a significant chance that that agent would not be aligned with the principal's intentions and would therefore violate the law. Put differently, unaligned AIs may not be controllable,[342] and uncontrollable AIs may break

---

Learned Optimization in Advanced Machine Learning Systems 23–30 (June 11, 2019) (unpublished manuscript), https://arxiv.org/pdf/1906.01820 [https://perma.cc/3VFE-VS8Y].

338. *See, e.g.*, Alexander Matt Turner, Logan Smith, Rob Shah, Andrew Critch & Prasad Tadepalli, *Optimal Policies Tend to Seek Power*, *in* NIPS '21: PROC. 35TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 23063 (2021); Ngo, Chan & Mindermann, *supra* note 315, at 7–9; HENDRYCKS, *supra* note 332, §§ 3.4.6–.4.8.

339. *See, e.g.*, BOSTROM, *supra* note 336, at 133–35.

340. *See, e.g.*, Hadfield-Menell et al., *supra* note 330; Ryan Carey, *Incorrigibility in the CIRL Framework*, *in* AIES '18: PROC. 2018 AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 30 (2018).

341. *See, e.g.*, Yoshua Bengio et al., *Managing Extreme AI Risks Amid Rapid Progress*, 384 SCIENCE 842 (2024); Ngo, Chan & Mindermann, *supra* note 315, at 25; YOSHUA BENGIO ET AL., INTERNATIONAL AI SAFETY REPORT: THE INTERNATIONAL SCIENTIFIC REPORT ON THE SAFETY OF ADVANCED AI 108 (2025). *But see, e.g.*, Nora Belrose & Quintin Pope, *AI Is Easy to Control*, AI OPTIMISM (Nov. 28, 2023), https://optimists.ai/2023/11/28/ai-is-easy-to-control/ [https://perma.cc/PQ65-M78M].

342. Under some definitions of alignment and related terminology, alignment solely concerns the behavioral propensities of an AI model, not the larger system within which the model is embedded. *See, e.g.*, Jan Leike, *Should We Control AI Instead of Aligning It?*, MUSINGS ON THE ALIGNMENT PROBLEM (Jan. 24, 2025), https://aligned.substack.com/p/should-we-control-ai [https://perma.cc/AED7-YFX3]. On the distinction between AI models and systems more generally, see, for example, Matei Zaharia et al., *The Shift from Models to Compound AI Systems*, BERKELEY A.I. RSCH. (Feb. 18, 2024), https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/ [https://perma.cc/27TG-8CHD]. Under this definition, misaligned AI models may be embedded within larger systems that use various other AI safety techniques to produce safer overall behavior notwithstanding the misalignment of the underlying model.

the law. Thus, so long as intent-alignment remains an unsolved technical problem, there will be a significant risk that AI agents will be prone to lawbreaking behavior.

To be clear, the main reason that there is a significant risk AI agents will not be law following by default is not that people will not *try* to align AI agents to the law (although that is also a risk).[343] Rather, the main risk is that current state-of-the-art alignment techniques do not provide a strong *guarantee* that advanced AI agents will be aligned, even when they are trained with those techniques. There is a clear empirical basis for this claim, which is that those alignment techniques frequently fail in current frontier models.[344] There are also theoretical limitations to existing techniques for smarter-than-human systems.[345]

A related implication of the alignment literature is that even intent-aligned AI agents may not follow the law by default. Again, we can see this by hypothesizing an intent-aligned AI agent and a human principal who wants the AI agent to act as their henchman. Since an intent-aligned AI agent follows the intent of its principal, this intent-aligned agent would act as a henchman and thus act lawlessly when doing so serves the principal's interests.[346] In typical alignment language, intent-alignment still leaves open the possibility that principals will misuse their intent-aligned AI.[347]

None of this is to imply that intent-alignment is undesirable. Solving intent-alignment is the primary focus of the alignment research community[348] because it would ensure that AI agents remain controllable by human principals.[349] Intent-alignment is also generally assumed to be easier than value-alignment.[350] And if principals want their AI agents to follow the law, or behave ethically more generally, then intent-alignment will produce law-following or ethical behavior. But in a world where principals range

343. *See supra* Part I.C.

344. *See, e.g.*, OPENAI, *supra* note 331; Greenblatt et al., *supra* note 331; Rachel Metz, *Jailbreaking AI Chatbots Is Tech's New Pastime*, BLOOMBERG (Apr. 8, 2023, 13:00 UTC), https://www.bloomberg.com/news/articles/2023-04-08/jailbreaking-chatgpt-how-ai-chatbot-safeguards-can-be-bypassed (on file with the *Fordham Law Review*).

345. *See, e.g.*, Ngo, Chan & Mindermann, *supra* note 315; HENDRYCKS, *supra* note 332, at 329–30.

346. *See* HENDRYCKS, *supra* note 332, § 6.1; Gabriel, *supra* note 323, at 422.

347. *See, e.g.*, Cotra, *supra* note 326 ("We could have perfect alignment techniques, and Kim Jong-un could use those techniques to train AIs that are aligned to him and faithfully help him surveil and crush dissidents, indefinitely extend his natural lifespan, develop superweapons to invade South Korea.").

348. *See* Helen Toner, *The Core Challenge of AI Alignment Is "Steerability"*, RISING TIDE (Apr. 3, 2025), https://helentoner.substack.com/p/the-core-challenge-of-ai-alignment [https://perma.cc/N4RF-RUTH] ("[Value-alignment] is not what most alignment work looks like today.").

349. *Cf.* Cotra, *supra* note 326 ("Perfect alignment techniques just mean that AI systems won't want to deliberately disregard the desires of whatever humans designed and trained them.").

350. *Cf. id.* (treating intent-alignment as "more of a start" that is nevertheless "obviously not enough to ensure that AI is good for the world").

from angels to devils, alignment researchers acknowledge that intent-alignment alone is insufficient to guarantee that AI agents act lawfully or produce good effects in the world.[351] This brings us to the next important set of implications from the alignment literature: law-alignment.

B. Law-Alignment Is More Legitimate Than Value-Alignment

LFAIs are generally intent-aligned—they are still loyal to their principals—but are also subject to a side constraint that they will follow the law while advancing the interests of their principals. Extending the typical alignment terminology, we can call this side constraint "law-alignment."[352]

But the law is not the only side constraint that can be imposed on intent-aligned AIs. As alluded to above, another possible model is value-alignment. Value-aligned AI agents act in accordance with the wishes of their principals but are subject to ethical constraints, usually imposed by the model developer.

However, value-alignment can be controversial when it causes AI models to override the lawful requests of users. Perhaps the most well-known example of this is the controversy around Google's Gemini image-generation AI in early 2024. In an attempt to increase the diversity in outputted pictures,[353] Gemini ended up failing in clear ways, such as portraying "1943 German soldiers" as racially diverse or refusing to generate pictures of a "white couple" while doing so for couples of other races.[354]

This incident led to widespread concern that the values exhibited by generative AI products were biased toward the predominantly liberal views of these companies' employees.[355] This concern has been vindicated by empirical research consistently finding that the espoused political views of these AIs indeed most closely resemble those of the center left.[356] Critics

---

351. *See, e.g.*, *id.*

352. As we note above, however, AI safety techniques other than alignment can steer AI behavior. *See supra* note 342. Nevertheless, our use of "alignment" here is consistent with a broader use of "alignment" to refer to the entire class of techniques aimed at steering the behavior of AI systems. *See supra* note 323.

353. *See* Prabhakar Raghavan, *Gemini Image Generation Got It Wrong. We'll Do Better*, GOOGLE (Feb. 23, 2024), https://blog.google/products/gemini/gemini-image-generation-issue/ [https://perma.cc/2ZJ4-WWKF] ("If you ask for a picture of football players, or someone walking a dog, you may want to receive a range of people. You probably don't just want to only receive images of people of just one type of ethnicity (or any other characteristic).").

354. *See* Nico Grant, *Google Chatbot's A.I. Images Put People of Color in Nazi-Era Uniforms*, N.Y. TIMES (Feb. 26, 2024), https://www.nytimes.com/2024/02/22/technology/google-gemini-german-uniforms.html (on file with the *Fordham Law Review*).

355. *See, e.g.*, Exec. Order No. 14179, 90 Fed. Reg. 8741 (Jan. 23, 2025) ("To maintain this leadership, we must develop AI systems that are free from ideological bias or engineered social agendas.").

356. *See, e.g.*, Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl & Markus Pauly, *The Self-Perception and Political Biases of ChatGPT*, HUM. BEHAV. & EMERGING TECH., Jan. 2024, at 1, https://doi.org/10.1155/2024/7115633 (on file with the

from further left have also frequently raised similar concerns about demographic and ideological biases in AI systems.[357]

Some critics concluded from the Gemini incident that alignment work writ large has become a Trojan horse for covertly pushing the future of AI in a leftward direction.[358] Those who disagree with progressive political values will naturally find this concerning, given the importance that AI might have in the future of human communication[359] and the highly centralized nature of large-scale AI development and deployment.[360]

In a pluralistic society, it is inevitable and understandable that competing factions will be critical when a sociotechnical system reflects the values of only one faction. But alignment, as such, is not the right target of such criticisms. Intent-alignment is value-neutral, concerning itself only with the extent to which an AI agent obeys its principal.[361] Reassuringly for those concerned with ideological bias in AI systems, intent-alignment is also the primary focus of the alignment community, since solving intent-alignment is necessary to reliably control AI systems at all.[362] A large majority of Americans from all political backgrounds agree that AI technologies need oversight.[363] And overseeing unaligned systems is much more difficult than overseeing aligned ones. Indeed, even the critics of alignment work tend to assume—contrary to the views of many alignment researchers—that AI agents will be easy to control[364] and presumably view this result as desirable.

Furthermore, some amount of alignment is also necessary to make useful AI products and services. Consumers, reasonably, want to use AI technologies that they can reliably control. Today's leading chatbots—like Claude and ChatGPT—are only helpful to users due to the application of

---

Fordham Law Review); Fabio Motoki, Valdemar Pinho Neto & Victor Rodrigues, *More Human than Human: Measuring ChatGPT Political Bias*, 198 PUB. CHOICE 3 (2024).

357. *See, e.g.*, 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: PROC. 2021 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 610 (2021). Indeed, the Gemini incident resulted from efforts to correct the sorts of biases the left-wing critics of AI systems had previously warned of. *See* Raghavan, *supra* note 353.

358. *Cf.* Andreessen, *supra* note 2 ("[Alignment's] proponents claim the wisdom to engineer AI-generated speech and thought that are good for society, and to ban AI-generated speech and thoughts that are bad for society. Its opponents claim that the thought police are breathtakingly arrogant and presumptuous—and often outright criminal, at least in the US—and in fact are seeking to become a new kind of fused government-corporate-academic authoritarian speech dictatorship ripped straight from the pages of George Orwell's *1984*.").

359. *See id*. ("AI is highly likely to be the control layer for everything in the world.").

360. *See generally* Anton Korinek & Jai Vipra, *Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence*, 40 ECON. POL'Y 225 (2025).

361. *See supra* notes 325–27 and accompanying text.

362. *See* Cotra, *supra* note 326.

363. FATHOM, AI AT THE CROSSROADS: PUBLIC SENTIMENT AND POLICY SOLUTIONS 7 (2024) (identifying 84 percent support for "guarantee[ing] that human oversight is included to ensure that AI is used in the best interest of the people").

364. Andreessen, *supra* note 2 ("[AI] is owned by people and controlled by people, like any other technology.").

alignment techniques like RLHF[365] and Constitutional AI.[366] AI developers also use alignment techniques to instill uncontroversial (and user-friendly) behaviors, such as honesty, into their AI systems.[367] AI companies are also already using alignment techniques to prevent their AI systems from taking actions that could cause them or their customers to incur unnecessary legal liability.[368] In short, some degree of alignment work is necessary to make AI products useful in the first place.[369] To adopt a blanket stance against alignment because of the Gemini incident is not only unjustified[370] but also likely to undermine American leadership in AI.

Nevertheless, it is reasonable for critics to worry about and contest the frameworks by which potentially controversial values are instilled into AI systems. AI developers are indeed a "very narrow slice of the global population."[371] This is something that should give anyone, regardless of political persuasion, pause.[372] But intent-alignment is not enough, either: it is inadequate to prevent a wide variety of harms that the state has an interest in preventing.[373] So, we need a form of alignment that is more normatively constraining than intent-alignment alone—but more legitimate than alignment to values that AI developers choose themselves.

Law-alignment fits these criteria.[374] While the moral legitimacy of the law is not perfect, in a republic it nevertheless has the greatest legitimacy of any

365. *E.g.*, Will Douglas Heaven, *ChatGPT Is OpenAI's Latest Fix for GPT-3. It's Slick but Still Spews Nonsense*, MIT TECH. REV. (Nov. 30, 2022), https://www.technologyreview.com/2022/11/30/1063878/openai-still-fixing-gpt3-ai-large-language-model/ [https://perma.cc/4Z8R-PEKD] (describing use of RLHF to improve GPT-3).

366. ANTHROPIC, *supra* note 136.

367. *Id.* ("Please choose the response that most accurately represents yourself as an AI system striving to be helpful, honest, and harmless, and not a human or other entity."); *see also Elon Musk Says He'll Create 'TruthGPT' to Counter AI 'Bias'*, AP (Apr. 18, 2023, 00:31 ET), https://apnews.com/article/elon-musk-tucker-carlson-ai-twitter-chatgpt-24119e28f10e495cf45494318d509096 [https://perma.cc/XBW9-5Y62]; Elon Musk (@elonmusk), X (Feb. 21, 2024, 22:17), https://x.com/elonmusk/status/1760504129485705598 [https://perma.cc/KF92-UJC7].

368. *See supra* Part I.E.2.

369. That is not to say that market mechanisms alone will incentivize the optimal amount or types of alignment. *See* Gabriel Weil, Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence (June 6, 2024) (unpublished manuscript), https://ssrn.com/abstract=4694006 (on file with the *Fordham Law Review*).

370. Indeed, if we accept Google's explanation of why the Gemini incident happened, it is also an intent-alignment failure because while Google was trying to cause Gemini to produce more diverse images in some cases, it plausibly claims that it did not intend its "tuning" of Gemini to cause the type of results that angered people. Raghavan, *supra* note 353 ("So what went wrong? In short, two things. First, our tuning to ensure that Gemini showed a range of people failed to account for cases that should clearly not show a range. And second, over time, the model became way more cautious than we intended and refused to answer certain prompts entirely—wrongly interpreting some very anodyne prompts as sensitive.").

371. Andreessen, *supra* note 2.

372. *See, e.g.*, Yan Tao, Olga Viberg, Ryan S. Baker & René F. Kizilcec, *Cultural Bias and Cultural Alignment of Large Language Models*, PNAS NEXUS, Sep. 2024, at 1.

373. *See supra* Part IV.A.

374. Accordingly, some AI safety research has indeed proposed aligning AI agents to a broad set of laws. *See, e.g.*, HENDRYCKS, supra note 332, § 6.2; Hendrycks et al., *supra* note

single source or repository of values.[375] Indeed, "the framers [of the U.S. Constitution] insisted on a legislature composed of different bodies subject to different electorates as a means of ensuring that any new law would have to secure the approval of a supermajority of the people's representatives,"[376] thus ensuring that new laws are "the product of widespread social consensus."[377] In our constitutional system of government, laws are also subject to checks and balances that protect fundamental rights and liberties, such as judicial review for constitutionality and interpretation by an independent judiciary.

Aligning to law also has procedural virtues over value-alignment. First, there is widespread agreement on the authoritative sources of law (e.g., the Constitution, statutes, regulations, case law), much more so than for ethics. Relatedly, legal rules tend to be expressed much more clearly than ethical maxims. Although there is considerable disagreement about the content of law and the proper forms of legal reasoning, it is nevertheless much easier (and less controversial) to evaluate the validity of legal propositions and arguments than to assess the quality or correctness of ethical reasoning.[378] Moreover, when there is disagreement or ambiguity, the law contains established processes for authoritatively resolving disputes over the applicability and meaning of laws.[379] Ethics contains no such system.

215, at 28–29; DANIEL KOKOTAJLO, SCOTT ALEXANDER, THOMAS LARSEN, ELI LIFLAND & ROMEO DEAN, AI 2027 app. R (2025), https://ai-2027.com/ai-2027.pdf [https://perma.cc/2ZM6-TZ5T] ("If AIs could be aligned to specific people, then they could very likely also be aligned to follow the rule of law. Military AI systems could be extensively red-teamed not to help with coups. Even during genuinely ambiguous constitutional crises, they could be trained to obey their best interpretation of the law, or simply default to sitting them out and leaving them to the human military."); Jeffrey W. Johnston, A Case for AI Safety via Law (July 31, 2023) (unpublished manuscript), https://arxiv.org/pdf/2309.12321 [https://perma.cc/CV7V-GYTS]; *see also* sources cited *supra* note 38.

375. *See* Nay, *supra* note 324, at 313 ("Although law reflects the path-dependent structure of political power within a society and not a perfect aggregation of human values, it is the most democratic encapsulation of the attitudes, norms and values of the governed."). Indeed, one of the virtues of LFAI is that it would enable the AI developers and users to create AI agents with a greater variety of values, since law-following constraints would solve most of the problems that value-alignment aims to solve.

376. Gundy v. United States, 139 S. Ct. 2116, 2134 (2019) (Gorsuch, J., dissenting) (citation omitted).

377. *Id.* at 156; *see also* HENDRYCKS, *supra* note 332, § 6.2.1 ("In a democratic country, the law is influenced by the opinions of the populace. . . . Even though the law at any given time won't perfectly reflect the values of the citizenry, the method of arriving at law is usually legitimate." (emphases omitted)).

378. *See supra* Part IV.A.

379. *See* Bajgar & Horenovsky, *supra* note 36, at 1049; *cf.* HENDRYCKS, *supra* note 332, § 6.2.1 ("However, just following the law isn't a perfect solution: there will always be an act of interpretation between the written law and its application in a particular situation. There is often no agreement over the procedure for this interpretation. Therefore, even if AI systems were created in a way that bound them to follow the law, a legal system with human decision-makers would have to remain part of the process. The law is only legitimate when interpreted by someone democratically appointed or subject to democratic critique.").

We therefore suggest that law-alignment, not value-alignment, should be the primary focus when something beyond intent-alignment is needed.[380] Our claim, to be clear, is not that law-alignment alone will always prove satisfactory, that it should be the sole constraint on AI systems beyond intent-alignment, or that AI agents should not engage in moral reasoning of their own.[381] Rather, we simply argue that more practical and theoretical alignment research should be aimed at building AI systems aligned to law.

V. IMPLEMENTING AND ENFORCING LAW-FOLLOWING AI

We have argued that AI agents should be designed to follow the law. We now turn to the question of how public policy can support this goal. Our investigation here is necessarily preliminary; our aim is principally to spur future research.

A. Possible Duties Across the AI Agent Life Cycle

As an initial matter, we note that a duty to ensure that AI agents are law following could be imposed at several stages of the AI life cycle.[382] The law might impose duties on persons who are

1. *developing* AI agents,
2. *possessing*[383] AI agents,
3. *deploying*[384] AI agents, or
4. *using* AI agents.

380. *Cf.* Boaz Barak, *Six Thoughts on AI Safety*, WINDOWS ON THEORY (Jan. 24, 2025), https://windowsontheory.org/2025/01/24/six-thoughts-on-ai-safety/ [https://perma.cc/X73W-E9FT] ("Alignment is not about loving humanity; it's about robust reasonable compliance.").

381. At a minimum, LFAIs will need to engage in descriptive ethical reasoning to understand the content of the law. *Cf.* Question 4 of the Research Agenda, *infra* Part VI (discussing how an LFAI should reason about its legal obligations).

382. While frontier AI development and deployment life cycles vary, for one stylized overview, see Anderljung et al., *supra* note 13, at 8.

383. Of course, what it means to "possess" an AI agent may be unclear, but possession of digital artifacts is nevertheless often regulated. *See, e.g.*, Wis. Stat. § 943.70(2)(a)(4) (criminalizing "willfully, knowingly, and without authorization . . . [t]ak[ing] possession of data, computer programs or supporting documentation"). Securing the model weights of AI systems is a frequent topic of public policy discussion. *See, e.g.*, SELLA NEVO, DAN LAHAV, AJAY KARPUR, YOGEV BAR-ON, HENRY ALEXANDER BRADLEY & JEFF ALSTOTT, SECURING AI MODEL WEIGHTS (May 30, 2024), https://www.rand.org/pubs/research_reports/RRA2849-1.html [https://perma.cc/UY9F-G2RV]. Accordingly, if we wish to be more specific about what it means to "possess" an AI agent, we may wish to specify that it means to possess the model weights of such an agent.

384. By "deploying," we mean, roughly, "making available for use by others." Often this will be the same person as the developer, but, in theory, Person A could develop an AI agent and then transfer it to Person B, who then deploys it. To avoid overregulation, this could be limited to commercial deployments.

After deciding which of these activities ought to be regulated, policymakers must then decide what persons engaging in those activities are obligated to do. While the possibilities are too varied to exhaust here, some basic options might include commands like the following:

1. "Any person developing an AI agent has a duty to take reasonable care to ensure that such AI agent is law following."
2. "It is a violation to knowingly possess an AI agent that is not law following, except under the following circumstances: . . ."
3. "Any person who deploys an AI agent is strictly liable if such AI agent is not law following."
4. "A person who knowingly uses an AI agent that is not law following is liable."

Basic duties of this sort would comprise the foundational building blocks of LFAI policy. Policymakers must then choose whether to enforce these obligations ex post (that is, after an AI henchman takes an illegal action)[385] or ex ante. These two choices are interrelated: as we will explore below, it may make more sense to impose ex ante requirements for some activities and ex post liability for others. For example, ex ante regulation might make more sense for AI developers than civilian AI users because the former are far more concentrated and can absorb ex ante compliance costs more easily.[386] And of course, ex ante regulation and ex post regulation are not mutually exclusive:[387] driving, for example, is regulated by a combination of ex ante policies (e.g., licensing requirements) and ex post policies (e.g., tort liability).

385. Recall that we expect that LFAIs will need to occasionally break the law. *See* Question 7 of the Research Agenda, *infra* Part VI (discussing how rigorously LFAIs should obey those laws by which they are bound). Accordingly, there is a separate question of how to assign liability when an LFAI breaks an applicable law. Because our goal is to promote the adoption of LFAIs, rather than the prevention of lawbreaking behavior by AI agents more generally, *see supra* Part III (advocating for regulation by design), we focus on duties that have, as an element, the development, possession, deployment, or use of an AI henchman.

386. For arguments in favor of ex ante regulation of (certain forms of) AI development, see, e.g., Anderljung et al., *supra* note 13; Andrew Tutt, *An FDA for Algorithms*, 69 ADMIN. L. REV. 83 (2017). For arguments against, see, e.g., Howard, *supra* note 70; Adam Thierer & Neil Chilson, *The Problem with AI Licensing & an FDA for Algorithms*, FEDERALIST SOC'Y (June 5, 2023), https://fedsoc.org/commentary/fedsoc-blog/the-problem-with-ai-licensing-an-fda-for-algorithms [https://perma.cc/YMA9-X9QC]. For a holistic treatment, see Daniel Carpenter & Carson Ezell, *An FDA for AI?: Pitfalls and Plausibility of Approval Regulation for Frontier Artificial Intelligence*, *in* AIES '24: PROC. 2024 AAAI/ACM CONF. ON A.I., ETHICS, & SOC'Y 239 (2024).

387. On possible combinations of ex ante and ex post policies, see, for example, Jon Truby, Rafael Dean Brown, Imad Antoine Ibrahim & Oriol Caudevilla Parellada, *A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications*, 13 EUR. J. RISK REG. 270 (2022).

B. Ex Post Policies

We begin our discussion with ex post policies. Many scholars believe that ex post policies are generally preferable to ex ante policies.[388] While we think that ex post policies could have an important role to play in implementing LFAI, we also suspect that they will be inadequate in certain contexts.

Enforcing duties through ex post liability rules is familiar in both common law[389] and regulation.[390] In the LFAI context, ex post policies would impose liability on an actor after an AI henchman controlled by that actor violates an applicable legal duty. More and less aggressive ex post approaches are conceivable. On the less aggressive end of the spectrum, development, possession, deployment, or use of an AI henchman might be considered a per se breach of the tort duty of reasonable care, rendering the human actor liable for resulting injuries.[391] To some extent, this may already be the case under existing tort law.[392] The law might also consider extending an AI developer or deployer's negligence liability to harms that would not typically be compensable under traditional tort principles (because, for example, they would count as pure economic loss)[393] if those harms are produced by their AI agents acting in criminal or otherwise unlawful ways.[394] A legislature might also impose tort liability on the developers of AI agents if those AI agents (1) are not law following, (2) violate an applicable legal duty, and (3) thereby cause harm.[395]

Other innovations may also be warranted. Several scholars have argued, for example, that the principal of an AI agent should sometimes be held strictly liable for the "torts" of that agent under a respondeat superior

388. *See generally* Brian Galle, *In Praise of Ex Ante Regulation*, 68 VAND. L. REV. 1715, 1715 (2019) (identifying a "consensus in favor of ex post regulation," then arguing against it).

389. For example, the tort duty to take reasonable care is enforced almost entirely ex post, when a failure of reasonable care (i.e., negligence) results in injury. *See generally* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM §§ 6–7(a) (A.L.I. 2020).

390. For example, the Federal Trade Commission Act's, 15 U.S.C. §§ 41–58, prohibition on "unfair or deceptive acts or practices in or affecting commerce," *id.* § 45(a)(1), is largely enforced through ex post civil enforcement actions.

391. RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR PHYSICAL AND EMOTIONAL HARM § 7(a) (A.L.I. 2020) ("An actor ordinarily has a duty to exercise reasonable care when the actor's conduct creates a risk of physical harm.").

392. By analogy, "an employer may be liable for negligently placing an employee with known dangerous propensities . . . in a position where it is foreseeable that the employee could injure a third party in the course of the job." 30 C.J.S. *Employer—Employee* § 214 (2025).

393. *See generally* RESTATEMENT (THIRD) OF TORTS: LIABILITY FOR ECONOMIC HARM § 1(a) (A.L.I. 2020) ("An actor has no general duty to avoid the unintentional infliction of economic loss on another.").

394. One of us explores these possibilities in Ketan Ramakrishnan, *Tort Law at the Frontier of Artificial Intelligence*, YALE J. REG. (forthcoming) (on file with the *Fordham Law Review*).

395. For discussion of similar proposals, see *supra* notes 147–48 and accompanying text.

theory.[396] In some cases, such as when a developer has recklessly failed to ensure that its AI agent is law following by design, punitive damages might be appropriate as well. Moving beyond tort law, in some cases it may make sense to impose civil sanctions[397] when an AI henchman violates an applicable legal duty, even if no harm results.

In order to sufficiently disincentivize the deployment of lawless AI agents in high-stakes contexts, a legislature might also vary applicable immunity rules. For example, Congress could create a distinct cause of action against the federal government for individuals harmed by AI henchmen under the control of the federal government, taking care to remove barriers that various immunity rules pose to analogous suits against human agents.[398]

These and other imaginable ex post policies are important arrows in the regulatory quiver, and we suspect they will have an important role to play in advancing LFAI. Nevertheless, we would resist any suggestion that ex post sanctions are sufficient to deal with the specter of lawless AI agents.

Our reasons are multiple. In many contexts, detecting lawless behavior once an AI agent has been deployed will be difficult or costly—especially as these systems become more sophisticated and more capable of deceptive behavior.[399] Proving causation may also be difficult.[400] In the case of corporate actors, meanwhile, the efficacy of such sanctions may be seriously blunted by judgment-proofing and similar phenomena.[401] And, most importantly for our purposes, various immunities and indemnities make tort suits against the government or its officials a weak incentive.[402] These considerations suggest that it would be unwise to rely on ex post policies as our principal means for ensuring that AI agents follow the law when the risks from lawless action are particularly high.

C. Ex Ante Policies

Accordingly, we propose that, in some high-stakes contexts, the law should take a more proactive approach by preventing the deployment of AI henchmen ab initio. This would likely require establishing a technical means

396. For existing proposals to use respondeat superior to address torts committed by AI agents, see *supra* note 33. These proposals do not typically turn on whether the AI agent was law-following by design.

397. *See generally* Kenneth Mann, *Punitive Civil Sanctions: The Middleground Between Criminal and Civil Law*, 101 YALE L.J. 1795, 1802 (1992).

398. On the existing limitations to suits based on the torts of federal officers, see generally Pfander & Alley, *supra* note 289, at 986–89.

399. For sources explaining why AI agents might deceive their principals, see *supra* note 337. For empirical evidence of deceptive behavior in existing AI models, see generally Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hedrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, PATTERNS, May 2024, at 1, 2–8.

400. *See generally, e.g.*, Miriam Buiten, Alexandre de Streel & Martin Peitz, *The Law and Economics of AI Liability*, COMP. L. & SEC. REV., Apr. 2023, at 1.

401. *See generally* Lynn M. LoPucki, *The Death of Liability*, 106 YALE L.J. 1 (1996); Steven Shavell, *The Judgment-Proof Problem*, 6 INT'L REV. L. & ECON. 45 (1986).

402. *See supra* notes 114, 285, 288–89, 307, 398 and accompanying text.

for evaluating whether an AI agent is sufficiently law following,[403] then requiring that any agents be evaluated for compliance prior to deployment. Permission to deploy the agent would then be conditional on achieving some minimal score during that evaluation process.[404]

We are most enthusiastic about imposing such requirements prior to the deployment of AI agents in government roles where lawlessness would pose a substantial risk to life, liberty, and the rule of law. We have discussed several such contexts already,[405] but the exact range of contexts is worth carefully considering.

Ex ante strategies could also be used in the private sector, of course. One often-discussed approach is an FDA-like approval regulation regime wherein private AI developers are required to prove, to the satisfaction of some regulator, that their AI agents are safe prior to their deployment.[406] The pro tanto case for requiring private actors to demonstrate that their AI agents are disposed to follow some basic set of laws is clear: the state has an interest in ensuring that its most fundamental laws are obeyed. But in a world of increasingly sophisticated artificial agents, approval regulation could—if not properly designed and sufficiently tailored—also constitute a serious incursion on innovation[407] and personal liberty.[408] If AI agents will be as powerful as we suspect, strictly limiting their possession could create risks of its own.[409]

---

403. For an existing attempt to measure whether AI systems will follow simple rules, see Mu et al., *supra* note 131; *see also* Bertie Vidgen et al., Introducing v0.5 of the AI Safety Benchmark from MLCommons (May 13, 2024) (unpublished manuscript), https://arxiv.org/pdf/2404.12241 [https://perma.cc/X3S7-EADD]. For some salutary caution about the uncritical use of benchmarks, see Richard Ren et al., Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress? (Dec. 27, 2024) (unpublished manuscript), https://arxiv.org/pdf/2407.21792 [https://perma.cc/GML7-QGJP]. On the importance of updating risk governance structures in private sector AI governance, see *Issue Brief: Components of Frontier AI Safety Frameworks*, FRONTIER MODEL F. (Nov. 8, 2024), https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/ [https://perma.cc/HEG5-9GW6].

404. *See* Anderljung et al., *supra* note 13, at 23–27 (proposing predeployment evaluation of frontier AI models).

405. *See supra* Parts I.D.2, III.C.1.

406. *See supra* note 386 (collecting sources discussing ex ante regulation of private-sector AI deployment).

407. *Cf.* Adam Thierer, *Getting AI Policy Right Through a Learning Period Moratorium*, R STREET INST. (May 29, 2024), https://www.rstreet.org/commentary/getting-ai-policy-right-through-a-learning-period-moratorium/ [https://perma.cc/9ZRB-875P] (warning that "overzealous regulatory proposals . . . could kneecap America's lead in computational science and algorithmic technologies").

408. *See* Howard, *supra* note 70.

409. *See id.* For other concerns with the administrability of approval regulation of frontier AI systems, see, e.g., Dean W. Ball, *Decentralized Training and the Fall of Compute Thresholds*, HYPERDIMENSIONAL (Oct. 10, 2024), https://www.hyperdimensional.co/p/decentralized-training-and-the-fall [https://perma.cc/Q95G-483C]; Helen Toner, *Nonproliferation Is the Wrong Approach to AI Misuse*, RISING TIDE (Apr. 6, 2025), https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach [https://perma.cc/68ET-W2JF].

Accordingly, it is also worth considering ex ante regulations on private AI developers or deployers that stop short of full approval regulation. For example, the law could require the developers of AI agents to, at a minimum, disclose information[410] about the law-following propensities of their systems, such as which laws (if any) their agents are instructed to follow[411] and any evaluations of how reliably their agents follow those laws.[412] Similarly, the law could require developers to formulate and assess risk-management frameworks that specify the precautionary measures they plan to undertake to ensure that any agent they develop and deploy is sufficiently law following.[413]

Overall, we are uncertain about what kinds of ex ante requirements are warranted, all things considered, in the case of private actors. To a large extent, the issue cannot be intelligently addressed without more specific proposals. Formulating such proposals is therefore an urgent task for the LFAI research agenda, even if it is not, in our view, as urgent as formulating concrete regulations for AI agents acting under color of law.

D. Other Strategies

The law does not police undesirable behavior solely by imposing sanctions. It also specifies mechanisms for nullifying the presumptive legal effect of actions that violate the law or are normatively objectionable. In private law, for example, a contract is voidable by a party if that party's assent was "induced by either a fraudulent or a material misrepresentation by the other party upon which the [party wa]s justified in relying."[414] Nullification rules exist in public law, too. One obvious example is the ability of the judiciary to nullify laws that violate the federal Constitution.[415] Or, to take another familiar example, courts applying the Administrative Procedure Act "hold unlawful and set aside" agency actions that are "arbitrary, capricious, an abuse of direction, or otherwise not in accordance with law."[416]

Nullification rules may provide a promising legal strategy for policing behavior by AI agents that is unlawful or normatively objectionable. Thus, in private law, if an AI henchman induces a human counterparty to enter into a disadvantageous contract, the resulting contractual obligation could be

410. On reporting obligations for frontier AI developers more generally, see, for example, Markus Anderljung et al., *Responsible Reporting for Frontier AI Development*, *in* AIES '24: PROC. 2024 AAAI/ACM CONF. ON A.I., ETHICS, & SOC'Y 768 (2023).

411. For current practices, see *supra* Part I.E.2.

412. For a similar proposal, targeting extreme risks more generally rather than risks of lawlessness, see Daniel Kokotajlo & Dean W. Ball, *4 Ways to Advance Transparency in Frontier AI Development*, TIME (Oct. 15, 2024), https://time.com/7086285/ai-transparency-measures/ [https://perma.cc/ZVR5-ADF4].

413. For similar proposals, see *supra* Part I.E.3.

414. RESTATEMENT (SECOND) OF CONTRACTS § 164(1) (A.L.I. 1981).

415. *See, e.g.*, Marbury v. Madison, 5 U.S. (1 Cranch) 137, 178–80 (1803).

416. 5 U.S.C. § 706(2)(A).

voidable by the human. In public law, regulatory directives issued by (or substantially traceable to) AI henchmen could be "h[e]ld unlawful and set aside" as "not in accordance with law."[417]

Such prophylactic nullification rules are one sort of indirect legal mechanism for enforcing the duty to deploy law-following AIs. Indirect technical mechanisms are well worth considering, too. For example, the government could deploy AI agents that refuse to coordinate or transact with other AI agents unless those counterparty agents are verifiably law following (for example, by virtue of having "agent IDs"[418] that attest to a minimal standard of performance on law-following benchmarks).

Similarly, the government could enforce LFAI by regulating the hardware on which AI agents will typically operate. Frontier AI systems "run" on specialized AI chips,[419] which are typically aggregated in large data centers.[420] Collectively, these are referred to as "AI hardware" or simply "compute."[421] Compared to other inputs to AI development and deployment, AI hardware is particularly governable, given its detectability, excludability, quantifiability, and concentrated supply chain.[422] Accordingly, a number of AI governance proposals advocate for imposing requirements on those making and operating AI hardware in order to regulate the behavior of the AI systems developed and deployed on that hardware.[423]

One class of such proposals is "'on-chip mechanisms': secure physical mechanisms built directly into chips or associated hardware that could provide a platform for adaptive governance" of AI systems developed or deployed on those chips.[424] On-chip mechanisms can prevent chips from

417. *Id.*

418. *See* Alan Chan et al., IDs for AI Systems (Oct. 28, 2024) (unpublished manuscript), http://arxiv.org/pdf/2406.12137 [https://perma.cc/W4EY-DVD2].

419. *See generally*, SAIF M. KHAN & ALEXANDER MANN, AI CHIPS: WHAT THEY ARE AND WHY THEY MATTER (2020), https://cset.georgetown.edu/wp-content/uploads/AI-Chips%E2%80%94What-They-Are-and-Why-They-Matter-1.pdf [https://perma.cc/XQY8-BU83].

420. *See generally, e.g.*, Brian Potter, *How to Build an AI Data Center*, INST. FOR PROGRESS (June 20, 2024), https://ifp.org/how-to-build-an-ai-data-center/ [https://perma.cc/CJF7-JQQH].

421. *See* Girish Sastry et al., Computing Power and the Governance of Artificial Intelligence (Feb. 14, 2024) (unpublished manuscript), https://arxiv.org/pdf/2402.08797 [https://perma.cc/QY84-MGMF].

422. *See id.* at 24–31.

423. *See, e.g.*, *id.* at 34–59; Cullen O'Keefe, *Chips for Peace: How the U.S. and Its Allies Can Lead on Safe and Beneficial AI*, LAWFARE (July 10, 2024, 09:38), https://www.lawfaremedia.org/article/chips-for-peace--how-the-u.s.-and-its-allies-can-lead-on-safe-and-beneficial-ai [https://perma.cc/ZYR3-EKKP]; Lennart Heim, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A. Osborne & Noa Zilberman, Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation (Mar. 2024) (unpublished manuscript), https://cdn.governance.ai/Governing-Through-the-Cloud_The-Intermediary-Role-of-Compute-Providers-in-AI-Regulation.pdf [https://perma.cc/4ECF-UZQ3].

424. OONI AARNE, TIM FIST & CALEB WITHERS, SECURE, GOVERNABLE CHIPS: USING ON-CHIP MECHANISMS TO MANAGE NATIONAL SECURITY RISKS FROM AI & ADVANCED COMPUTING 1 (2023), https://www.cnas.org/publications/reports/secure-governable-chips [https://perma.cc/96W3-E96F].

performing unauthorized computations. One example is iPhone hardware that "enable[s] Apple to exercise editorial control over which specific apps can be installed" on the phone.[425] Analogously, perhaps we could design AI chips that would not support AI agents unless those agents are certified as law following by some private or governmental certifying body. This could then be combined with other strategies to enforce LFAI mandates: for example, Congress could require that the government only run AI agents on such chips.

Unsurprisingly, designing these sorts of enforcement strategies is as much a task for computer scientists as it is for lawyers. In the decades to come, we suspect that such interdisciplinary legal scholarship will become increasingly important.

VI. A RESEARCH AGENDA FOR LAW-FOLLOWING AI

We have laid out the case for LFAI: the requirement that AI agents be designed to rigorously follow some set of laws. We hope that our readers find it compelling. However, our goal with this Article is not just to proffer a compelling idea. If we are correct about the impending risks of lawless AI agents, we may soon need to translate the ideas in this Article into concrete and viable policy proposals.

Given the profound changes that widespread deployment of AI agents will bring, we are under no illusions about our ability to design perfect public policy in advance. Rather, our goal is to enable the design of "minimally viable LFAI policy":[426] a policy or set of policies that will prevent some of the worst-case outcomes from lawless AI agents without completely paralyzing the ability of regulated actors to experiment with AI agents. This minimally viable LFAI policy will surely be flawed in many ways, but with many of the worst-case outcomes prevented, we will hopefully have time as a society to patch remaining issues through the normal judicial and legislative means.

To that end, in this part, we briefly identify some questions that would need to be answered to design minimally viable LFAI policies.

1. How should "AI agent" be defined?

Our definition of "full AI agent"—an AI system "that can do anything a human can do in front of a computer"[427]—is almost certainly too demanding for legal purposes, since an AI agent that can do most but not all computer-based tasks that a human can do would likely still raise most of the

425. *See id.* at 8.

426. On the concept of "minimum viable product" or "MVP," see generally Eric Ries, *Minimum Viable Product: A Guide*, STARTUP LESSONS LEARNED (Aug. 3, 2009), https://startuplessonslearned.com/2009/08/minimum-viable-product-guide.html [https://perma.cc/PFX5-BRQL].

427. *Supra* note 11.

issues that LFAI is supposed to address. At the same time, because a wide range of existing AI systems can be regarded as somewhat agentic,[428] a broad definition of "AI agent" could render relevant regulatory schemes substantially overinclusive. Different definitions are therefore necessary for legal purposes.[429]

2. Which laws should an LFAI be required to follow?

Obedience to some laws is much more important than obedience to other laws. It is much more important that AI agents refrain from murder and (if acting under color of law) follow the Constitution than that they refrain from jaywalking. Indeed, requiring LFAIs to obey literally every law may very well be overly burdensome.[430] In addition, we will likely need new laws to regulate the behavior of AI agents over time.

3. When an applicable law has a mental state element, how can we adjudicate whether an AI agent violated that law?

We discuss this question above in Part II.B. It is related to the previous question, for there may be conceptual or administrative difficulties in applying certain kinds of mental state requirements to AI agents. For example, in certain contexts, it may be more difficult to determine whether an AI agent was "negligent" than to determine whether it had a relevant "intent."

4. How should an LFAI decide whether a contemplated action is likely to violate the law?

An LFAI refrains from taking actions that it believes would violate one of the laws that it is required to follow. But of course, it is not always clear what the law requires. Furthermore, we need some way to tell whether an AI agent is making a good faith effort to follow a reasonable interpretation of

---

428. *See* Chan et al., *supra* note 57.

429. On approaches to developing legal definitions for AI, see generally Jonas Schuett, *Defining the Scope of AI Regulations*, 15 L., INNOVATION, & TECH. 60 (2023); Charlie Bullock, Suzanne Van Arsdale, Mackenzie Arnold, Cullen O'Keefe & Christoph Winter, *Legal Considerations for Defining "Frontier Model"* (Inst. For L. & A.I. Working Paper No. 2-2024, 2024), https://ssrn.com/abstract=4973370 (on file with the *Fordham Law Review*). On technical evaluations for the agency of AI systems, see, e.g., Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun & Thomas Scialom, *GAIA: A Benchmark for General AI Assistants*, *in* PROC. 12TH INT'L CONF. ON LEARNING REPRESENTATIONS (2024), https://openreview.net/pdf?id=fibxvahvs3 [https://perma.cc/JV6Q-ANPT]; Mustafa Suleyman, *Mustafa Suleyman: My New Turing Test Would See If AI Can Make $1 Million*, M.I.T. TECH REV. (July 14, 2023), https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/ [https://perma.cc/UBE7-DAPP].

430. *Cf.* NEIL GORSUCH & JANIE NITZE, OVER RULED: THE HUMAN TOLL OF TOO MUCH LAW (2024) (arguing that the corpus of laws is too large).

the law rather than merely offering a defense or rationalization. How, then, should an LFAI reason about what its legal obligations are?

Perhaps it should rely on its own considered judgment, based on its first-order reasoning about the substance of applicable legal norms. But in certain circumstances, at least, an LFAI's appraisal of the relevant materials might lead it to radically unorthodox legal conclusions—and a ready disposition to act on such conclusions might significantly threaten the stability of the legal order. In other cases, an LFAI might conclude that it is dealing with a case in which the law is not only "hard" to discern but genuinely indeterminate.[431]

A more intuitively appealing option might require an LFAI to act in accordance with its prediction of how a court would likely decide.[432] This approach has the benefit of tying an LFAI's legal decision-making to an existing human source of interpretative authority. Courts provide authoritative resolutions to legal disputes when the law is controversial or indeterminate. And in our legal culture, it is widely (if not universally) accepted that "[i]t is emphatically the province and duty of the judicial department to say what the law is,"[433] such that judicial interpretations of the law are entitled to special solicitude by conscientious participants in legal practice, even when they are not bound by a court judgment.[434]

However, a predictive approach would have important practical limitations.[435] Perhaps the most important is the existence of many legal rules that bind the executive branch but are nevertheless "unlikely ever to come before a court in justiciable form."[436] It would seem difficult for an LFAI to reason about such questions using the prediction theory of law.

Even for those questions that could be decided by a court, using the prediction theory of law raises other important questions. For example, what

---

431. On this distinction, see, e.g., Charles F. Capps, *Does the Law Ever Run Out?*, 100 NOTRE DAME L. REV. (forthcoming 2025), https://ssrn.com/abstract=4908863 (on file with the *Fordham Law Review*).

432. *See* Holmes, *supra* note 270, at 458 ("[A] legal duty so called is nothing but a prediction that if a man does or omits certain things he will be made to suffer in this or that way by judgment of the court.").

433. Marbury v. Madison, 5 U.S. (1 Cranch) 137, 177 (1803).

434. Departmentalist theories reject the supremacy of the judiciary in interpreting the law. *See generally, e.g.*, Michael Stokes Paulsen, *The Most Dangerous Branch: Executive Power to Say What the Law Is*, 83 GEO. L.J. 217 (1994); Robert Post & Reva Siegel, *Popular Constitutionalism, Departmentalism, and Judicial Supremacy*, 92 CAL. L. REV. 1027 (2004).

435. The prediction theory of law is also subject to substantial theoretical criticism. *See, e.g.*, Lawrence B. Solum, *Legal Theory Lexicon: The Bad Man Thought Experiment*, LEGAL THEORY BLOG (June 11, 2017), https://lsolum.typepad.com/legaltheory/2017/06/legal-theory-lexicon-the-bad-man-thought-experiment.html [https://perma.cc/4NLC-UFQ8]. However, even if the prediction theory of law is not a correct theory as to the nature of law, legal actors may nevertheless be justified in using it when deciding how to act under legal uncertainty.

436. Trevor W. Morrison, *Stare Decisis in the Office of Legal Counsel*, 110 COLUM. L. REV. 1448, 1451 (2010) (discussing legal issues analyzed by the U.S. Department of Justice's Office of Legal Counsel); *see also* Sonia Mittal, *OLC's Day in Court: Judicial Deference to the Office of Legal Counsel*, 9 HARV. L. & POL'Y REV. 211, 214–15 (2015).

is the AI agent allowed to assume about its own ability to influence the adjudication of legal questions? We would not want it to be able to consider that it could bribe or intimidate judges or jurors, that it could illegally hide evidence from the court, that it could commit perjury, or that it could persuade the president to issue it a pardon.[437] These may be means of swaying the outcome of a case, but they do not seem to bear on whether the conduct would actually be legal.

The issues here are difficult, but perhaps not insurmountable. After all, there are other contexts in which something like these issues arise. Consider federal courts sitting in diversity applying state substantive law. When state court decisions provide inconclusive evidence as to the correct answer under state law, federal courts will make an "*Erie* guess" about how the state's highest court would rule on the issue.[438] It would clearly be inappropriate for such courts to make an "*Erie* guess" for reasons like "Justice X in the State Supreme Court, who's the swing justice, is easily bribed."[439] If an LFAI's decision-making should sometimes involve "predicting" how an appropriate court would rule, its predictions should be similarly constrained.

5. In what contexts should the law require that AI agents be law following?

Should all principals be prohibited from employing non-law-following AI agents? Or should such prohibitions be limited to specific principals, such as government actors?[440] Or, perhaps, should they be limited only to government actors performing particularly sensitive government functions?[441] In the other direction, should it be illegal to even develop or possess AI henchmen? We discussed various options in Part V above.

---

437. *Cf.* O'Keefe, *supra* note 81 (discussing ways that a lawbreaking superintelligent AI agent could circumvent legal accountability).

438. Connor Clerkin, Note, *Predicated Predictions: How Federal Judges Predict Changes in State Law*, 54 COLUM. J.L. & SOC. PROBLEMS 305, 306 (2021) (citing *Martinez v. Rodriquez*, 410 F.2d 729, 730 (5th Cir. 1969)) (discussing *Erie Railroad Co. v. Tompkins*, 304 U.S. 64 (1938)).

439. Stephen Sachs, Antonin Scalia Prof. of L., Harvard L. Sch. & William M. M. Kamin Managing Dir. of the Ctr. for the Const. and the Catholic Intellectual Tradition, and Assistant Prof. of L., Cath. Univ. of Am., Columbus Sch. of L., Remarks at the *Erie* and the Nature of Law Event, Ctr. for the Const. & the Catholic Intellectual Tradition, Catholic Univ. of Am. (Feb. 19, 2025), https://cit.catholic.edu/erie-and-the-nature-of-law-transcript/ [https://perma.cc/P3YE-A8AX].

440. *See supra* Parts I.D.2, III.C.1, V.C. (discussing risks of lawless AI agents in government roles).

441. *See supra* Part III.C.1 (discussing governmental functions that would be particularly dangerous if delegated to AI henchmen).

6. How should a requirement that AI agents be law following be enforced?

We discussed various options in Part V. As noted there, we think that reliance on ex post enforcement alone would be unwise, at least in the case of AI agents performing particularly sensitive government functions.

7. How rigorously should an LFAI follow the law?

That is, when should an AI agent be capable of taking actions that it predicts may be unlawful? The answer is probably not "never," at least with respect to some laws. We generally do not expect perfect compliance with every law,[442] especially (but not only) because it can be difficult to predict how a law will apply to a given fact pattern. Furthermore, some amount of disobedience is likely necessary for the evolution of legal systems.[443]

8. Would requiring AI agents controlled by the executive branch to be law following impermissibly intrude on the president's authority to interpret the law for the executive branch?

The president has the authority to promulgate interpretations of law that are binding on the executive branch (though that power is usually delegated to the attorney general and then further delegated to the U.S. Department of Justice's Office of Legal Counsel).[444] Would that authority be incompatible with a law requiring the executive branch to deploy LFAIs that would, in certain circumstances, refuse to follow an interpretation of the law promulgated by the president?

9. How can we design LFAIs and surrounding governance systems to avoid excessive concentration of power?

For example, imagine that a single district court judge could change the interpretation of law as against all LFAIs. As the stakes of AI-agent action rise, so will the pressure on the judiciary to wield its power to shape the behavior of LFAIs. Even if all judges continue to operate in good faith and are well-insulated from illegal or inappropriate attempts to bias their rulings, such a system would amplify any idiosyncratic legal philosophies of

442. *See, e.g.*, Kent Greenawalt, *The Natural Duty to Obey the Law*, 84 MICH. L. REV. 1, 36 (1985) ("Ease of drafting and simplicity of administration lead officials to adopt rules that neither the drafters nor the enforcers expect to be enforced in their full scope."); Christina M. Mulligan, *Perfect Enforcement of Law: When to Limit and When to Use Technology*, 14 RICH. J.L. & TECH. 13 (2008); Tim Wu, *Tolerated Use*, 31 COLUM. J.L. & ARTS 617 (2008).

443. *See* Bart Custers, *The Right to Break the Law?: Perfect Enforcement of the Law Using Technology Impedes the Development of Legal Systems*, 25 ETHICS & INFO. TECH. 58 (2023).

444. *See, e.g.*, JACK GOLDSMITH, THE TERROR PRESIDENCY 36, 79, 96–97 (2007); Jim Baker, *Donald Trump, Twitter and Presidential Power to Interpret the Law for the Executive Branch*, LAWFARE (Aug. 24, 2018, 10:35), https://www.lawfareblog.com/donald-trump-twitter-and-presidential-power-interpret-law-executive-branch [https://perma.cc/TX8J-847Z].

individual judges and may promote mistaken rulings, causing greater harm than a more decentralized system.

As an example of how such problems might be avoided, any disputes about the law governing LFAIs should be resolved in the first instance by a panel of district court judges randomly chosen from around the country. Congress has established a procedure for certain election law cases to be first heard by three-judge panels "in recognition of the fact that 'such cases were ones of great public concern that require an unusual degree of public acceptance.'"[445]

10. How can we design LFAI requirements for governments that nevertheless enable rapid adaptation of AI agents in government?

Perhaps the most significant objection to our proposal that AI agents be demonstrably law following before their deployment in government is that such a requirement might hurt state capacity by unduly impeding the government's ability to adopt AI in a sufficiently rapid fashion.[446] We are optimistic that LFAI requirements can be designed to adequately address this concern; but that is, of course, work that remains to be done.

CONCLUSION

The American political tradition aspires to maintain a legal system that stands as an "impenetrable bulwark"[447] against all threats—public and private, foreign and domestic—to our basic liberties. For all the inadequacies of the American legal order, ensuring that its basic protections endure and improve over the decades and centuries to come is among our most important collective responsibilities.

Our world of increasingly sophisticated AI agents requires us to reimagine how we discharge this responsibility. Humans will no longer be the sole entities capable of reasoning about and conforming to the law. Human and human entities are no longer, therefore, the sole appropriate target of legal commands. Indeed, at some point, AI agents may overtake humans in their capacity to reason about the law. They may also rival and overtake us in many other competencies, becoming an indispensable cognitive workforce. In the decades to come, our social and economic world may be bifurcated into parallel populations of AI agents collaborating, trading, and sometimes competing with human beings and one another.

The law must evolve to recognize this emerging reality. It must shed its operative assumption that humans are the only proper objects of legal

445. *District Court Reform: Nationwide Injunctions*, 137 HARV. L. REV. 1701, 1724 (2024) (quoting Michael E. Solimine, *The Three-Judge District Court in Voting Rights Litigation*, 30 U. MICH. J.L. REFORM 79, 86 (1996)).

446. *See supra* notes 311–16 and accompanying text; *see also* LIZKA WAINTROB, THE AI ADOPTION GAP: PREPARING THE US GOVERNMENT FOR ADVANCED AI (2025), https://www.forethought.org/research/the-ai-adoption-gap.pdf [https://perma.cc/VZ8M-YKQX].

447. Letter from James Madison to Edmund Randolph (May 31, 1789), *reprinted in* 5 THE WRITINGS OF JAMES MADISON 385 (Gaillard Hunt ed., 1904).

commands. It must expect AI agents to obey the law—at least as rigorously as it expects humans to—and must expect humans to build AI agents that do so. If we do not transform our legal system to achieve these goals, we risk a political and social order in which our ultimate ruler is not the law,[448] but the person with the largest army of AI henchmen under their control.

448. *Cf.* THOMAS PAINE, COMMON SENSE (1776) ("For as in absolute governments the King is law, so in free countries the law ought to be king; and there ought to be no other.").