LAWFARE

LAWFARE RESEARCH REPORT 26-8

Proceedings of the 2025 Workshop on Law-Following AI

May 2026

The inaugural Workshop on Law-Following AI (LFAI), hosted by the Institute for Law & AI at the University of Cambridge from August 6–8, 2025 with support from the Leverhulme Centre for the Future of Intelligence and the UK's Advanced Research & Innovation Agency, convened more than forty scholars from law, computer science, and related disciplines to advance the emerging research agenda around LFAI: a concept denoting agentic AI systems designed to refuse illegal orders and illegal means, the corresponding policy proposal to mandate such design in certain deployment contexts, and the interdisciplinary field of inquiry that supports both. Rather than recording consensus, these proceedings synthesize key themes from the workshop's presentations and discussions, including the promises and limits of liability (particularly for governmental AI agents), the state and nuances of automated legal reasoning and evaluation, risks posed by automated compliance and "perfect enforcement," the appropriate standard of care for AI agents, and the interplay between AI agents and their principals, including fiduciary framings. The report is intended to extend the workshop's conversations to a broader audience and catalyze further scholarship on LFAI's design, evaluation, and governance. A full list of report authors can be found at the end of this document.

Table of Contents

Introduction

Law-following AI (LFAI)[1] denotes three closely related concepts. First, it denotes a plausibly desirable *form of AI technology*: agentic AI systems[2] designed to follow the law by refusing to follow illegal orders or use illegal means to accomplish lawful orders.[3] Second, it denotes an accompanying *policy proposal*: imposing legal requirements that agentic AI systems in certain settings be LFAIs.[4] Third, it refers to an emerging *field of inquiry* aimed at enabling the development of law-following AI agents and the implementation of LFAI policies.[5]

To catalyze further research into LFAI, the Institute for Law & AI, with support from the Leverhulme Centre for the Future of Intelligence at the University of Cambridge and the United Kingdom's Advanced Research & Innovation Agency, hosted the inaugural Workshop on Law-Following AI at the University of Cambridge. Held from August 6–8, 2025, the workshop convened more than forty scholars from law, computer science, and related disciplines to discuss LFAI and, hopefully, make progress on aspects of the LFAI research agenda. This proceedings report compiles some of the key insights from the sessions and discussions at the workshop.

Given the number and diversity of participants, individual authors or participants may disagree with many of the ideas or arguments presented herein. Rather than attempting to capture any particular consensus, this document generally attempts to summarize key ideas and discussions from the workshop, so that those who were unable

---

[1] *See generally* Cullen O'Keefe et al., *Law-Following AI: Designing Agents to Obey Human Laws*, 94 FORDHAM L. REV. 57 (2025) (introducing the concept). Sources that develop similar ideas include Umang Bhatt & Holli Sargeant, *When Should Algorithms Resign? A Proposal for AI Governance*, 57 COMPUTER 99 (2024); Elina Nerantzi & Giovanni Sartor, *'Hard AI Crime': The Deterrence Turn*, 44 OXFORD J. LEGAL STUD. 673 (2024); SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011). Subsequent work on LFAI includes Matthijs Maas & Tobi Olasunkanmi, *Treaty-Following AI* (Inst. for L. & A.I. Working Paper No. 1-2025, 2025), https://law-ai.org/treaty-following-ai/ [https://perma.cc/6VGW-NEXC]; Luxi He et al., Statutory Construction and Interpretation for Artificial Intelligence (Sept. 1, 2025) (unpublished manuscript), https://arxiv.org/abs/2509.01186; Katalina Hernandez Delgado, The Law-Following AI Framework: Legal Foundations and Technical Constraints: Legal Analogues for AI Actorship and Technical Feasibility of Law Alignment (Sep. 8, 2025) (unpublished manuscript), https://www.arxiv.org/abs/2509.08009.

[2] For background on AI agents in the context of law-following AI, see generally O'Keefe et al., *supra* note 1, at 66–69. For foundational materials on AI agency, see generally MICHAEL WOOLDRIDGE, AN INTRODUCTION TO MULTIAGENT SYSTEMS (2nd ed. 2009); Michael Wooldridge & Nicholas R. Jennings, *Intelligent Agents: Theory and Practice*, 10 KNOWLEDGE ENG'G REV. 115 (1995).

[3] *See, e.g.*, O'Keefe et al., *supra* note 1, at 62.

[4] *See id.* at 116–23.

[5] *See id.* at 58.

to attend can nevertheless benefit from them and continue the academic discussion on this topic.

The Case for Law-Following AI in Brief

While we encourage readers interested in the topic of LFAI to read the full article,[6] we summarize some of its key ideas here for readability.[7] LFAI is a set of propositions about *AI agents*. While the exact definition of AI agency remains debated, AI agents are generally AI systems that can "pursue more complex goals in more complex environments, exhibiting independent planning and adaptation to directly take actions in virtual or real-world environments."[8] The *Law-Following AI* article uses "AI systems 'that can do anything a human can do in front of a computer' as competently as a human expert"[9] as its definition for "AI agents." Importantly, this definition is *illustrative* only: a more formal (and likely broader) definition of AI agent would be necessary for actual policymaking.[10]

AI agents are often analyzed using familiar principal–agent frameworks.[11] A major goal of AI policy and safety research has been ensuring that AI agents are *intent-aligned*: that is, that the agent reliably acts as its principal intended.[12] Intent-alignment is an unsolved technical problem: AI developers currently do not know how to make AI systems that are intent-aligned.[13]

However, intent-alignment is likely inadequate in many ways. Principals might include bad actors, or actors that are simply indifferent to the interests of others. To

---

[6] *Id.*

[7] For more informal background on LFAI, see generally Cullen O'Keefe & Ketan Ramakrishnan, *AI Agents Must Follow the Law*, LAWFARE (May 14, 2025, at 07:00 ET), https://www.lawfaremedia.org/article/ai-agents-must-follow-the-law [https://perma.cc/3ALA-JFL7]; Kevin Frazier et al., *Lawfare Daily: Cullen O'Keefe on the Impending Wave of AI Agents*, LAWFARE (May 14, 2025, at 07:00 ET), https://www.lawfaremedia.org/article/lawfare-daily--cullen-o-keefe-on-the-impending-wave-of-ai-agents; *Should AI Agents Obey Human Laws? (with Cullen O'Keefe)*, FORECAST (Aug. 28, 2025), https://pnc.st/s/forecast/15605074/should-ai-agents-obey-human-laws-with-cullen-o-keefe-.

[8] HELEN TONER ET AL., THROUGH THE CHAT WINDOW AND INTO THE REAL WORLD: PREPARING FOR AI AGENTS 1 (2024) (emphasis omitted), https://doi.org/10.51593/20240034 [https://perma.cc/5FYE-SNQE].

[9] O'Keefe et al., *supra* note 1, at 60 (quoting *ACT-1: Transformer for Actions*, ADEPT (Sep. 14, 2022) (footnote omitted), https://www.adept.ai/blog/act-1 [https://perma.cc/K878-WZTX]).

[10] *See id.* at 123–24.

[11] *See id.* at 70.

[12] *See id.* at 108.

[13] *See id.* at 109.

address this class of problems, AI safety researchers have proposed *value-alignment*—wherein AI agents have extralegal normative constraints on their actions that principals cannot override—as a complement to intent-alignment.[14]

LFAI can be thought of as an alternative to value-alignment, and therefore a complement to intent-alignment.[15] While the law is certainly not a complete guide to moral behavior, aligning AI systems to the law has important advantages over value-alignment. One major advantage is *legitimacy*: democratically enacted laws are a much more legitimate source of constraints than extralegal moral principles.[16] Another advantage is *authoritativeness*. The law is generally drawn from a small set of well-specified authoritative sources, such as constitutions, statutes, and case law.[17] There is no comparable authoritative list of written moral principles that commands near-universal recognition. A related benefit is *precision*. While legal drafting is often vague or inartful, legal prohibitions tend to be much more clear about what, exactly, they require than moral injunctions.[18] Finally, the law is more easily *resolvable* than morality: "when there is disagreement or ambiguity, the law contains established processes for authoritatively resolving disputes over the applicability and meaning of laws."[19]

LFAI might be particularly important for AI agents controlled by governments.[20] LFAI is an ex ante means of preventing lawless government actions: LFAIs by definition *refuse to violate* applicable laws in the first place. Preventing lawless action by governmental AI agents ex ante is important for several reasons. First, ex post remedies, such as tort suits and criminal prosecutions for civil rights violations, play a limited role in constraining governmental abuse.[21] Immunity doctrines,[22] indemnification practices,[23] the prospect of pardons for criminal abuses, resource asymmetries, misaligned

[14] *See id.* at 108–09.

[15] *See id.* at 112–16.

[16] *See id.* at 112–15.

[17] *See id.* at 115.

[18] *See id.*

[19] *Id.*

[20] *See id.* at 98–99, 101–08, 119–21.

[21] For further discussion of the interplay between LFAI and liability, see *infra* section "The Promises and Limits of Liability."

[22] *See* O'Keefe et al., *supra* note 1, at 77 n.114, 105.

[23] *See* Joanna C. Schwartz, *Police Indemnification*, 89 N.Y.U. L. REV. 885, 890–91 (2014).

financial incentives,[24] and limited opportunities for self-help[25] all mean that ex post liability (whether criminal or civil) is a significantly weaker check against government agents than private actors. Accordingly, we largely use ex ante legal tools—injunctions,[26] nullification,[27] multiple independent veto points,[28] oaths, responsible hiring practices, supervision, and disqualification rules[29]—to *prevent* lawless governmental action in the first place. LFAI can be thought of as just such an ex ante tool: one that takes advantage of the designable nature of AI systems[30] to stop lawless action at its source.

Indeed, ex ante constraints might be more important for governmental AI agents than for human agents. Notwithstanding all of the limitations of ex post mechanisms, they still play a very important role in preventing lawless governmental action.[31] But AI agents that are merely intent-aligned would, by default, lack many of the incentives to follow the law that human agents have, such as fear of imprisonment,[32] reputational damage, and the unpleasantness of litigation or congressional oversight, not to mention innate personal morality. Ex ante law-following design can compensate for these ex post weaknesses.[33]

---

[24] Since settlements or damages are paid out of the public fisc, not the personal resources of the top executive branch officials, such officials may have relatively weak incentives to ensure that their subordinates do not commit torts.

[25] For example, many states have abolished the common-law right to resist unlawful arrests. *E.g.*, State v. Kutchara, 350 N.W.2d 924, 927 (Minn. 1984); State v. Thomas, 625 S.W.2d 115, 121–22 (Mo. 1981).

[26] Injunctions are the preferred remedy for vindication of constitutional rights. *See generally* John F. Preis, *In Defense of Implied Injunctive Relief in Constitutional Cases*, 22 WM. & MARY BILL RTS. J. 1 (2013); Andrew Kent, *Are Damages Different?:* Bivens *and National Security*, 87 S. CAL. L. REV. 1123 (2014).

[27] *See* O'Keefe et al., *supra* note 1, at 121–22.

[28] *See id.* at 98.

[29] *See id.* at 99.

[30] *See id.* at 93–108.

[31] *See* 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, at 12:52 ET), 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].

[32] *See* O'Keefe et al., *supra* note 1, at 102–03.

[33] Of course, it may still be valuable to pursue various ex post mechanisms for aligning the incentives of intent-aligned AI systems with broader societal values. *See, e.g.*, Peter Salib & Simon Goldstein, *AI Rights for Human Safety*, VA. L. REV. (forthcoming 2026), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4913167 (arguing that granting economic rights to AIs "would enable humans and AIs to engage in iterated, small-scale, mutually-beneficial transactions").

Major Themes from the Workshop

In this section, we summarize some of the major themes from presentations and discussions at the workshop. Since the goal of the workshop was to inspire further research on LFAI, many discussions took on a critical posture, such as identifying weaknesses or limitations in the LFAI article. The discussion herein reflects many of those critical themes. These critical discussions, however, should not be mistaken for an overall *evaluation* of LFAI; participants generally found the topic intellectually generative and meritorious.

The Promises and Limits of Liability

LFAI is hypothesized as a desirable complement to liability. In many cases where an AI agent behaves illegally and thereby causes harm, it will be possible to hold the agent's principal liable for that harm in tort. LFAI does not aim to remove the possibility of such actions.[34] However, as mentioned above,[35] LFAI is also premised on the assumption that tort will prove inadequate in many cases, such as where the principal is a government actor,[36] or where an AI system behaves much more culpably than its principal (for example, by using criminal means to satisfy an innocuous command), such that it would be unjust to hold the principal fully responsible.[37] Since LFAI is justified in part as a solution to the shortcomings of liability, an accurate appraisal of LFAI depends on an accurate understanding of how liability is likely to work—or fail—when AI agents are widely deployed.

The workshop contained lively discussion on these questions, drawing on a large and growing body of scholarly literature.[38] As with many discussions in technology

---

[34] *See* O'Keefe et al., *supra* note 1, at 63–64.

[35] *See supra* Part I.

[36] *See* O'Keefe et al., *supra* note 1, at 101–07, 120.

[37] This problem is known as the "responsibility gap." *See generally* Trystan S. Goetze, *Mind the Gap: Autonomous Systems, the Responsibility Gap, and Moral Entanglement*, FACCT '22: PROCS. 2022 ACM CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 390, https://doi.org/10.1145/3531146.3533106.

[38] For informative literature on this topic, see, for example, Mihailis E. Diamantis, *Reasonable AI: A Negligence Standard*, 78 VAND. L. REV. 573 (2025) [hereinafter Diamantis, *Reasonable AI*]; Mihailis E. Diamantis, *Vicarious Liability for AI*, 99 IND. L.J. 317 (2023) [hereinafter Diamantis, *Vicarious Liability*]; Mihailis E. Diamantis, *Employed Algorithms: A Labor Model of Corporate Liability for AI*, 72 DUKE L.J. 797 (2022); Mihailis E. Diamantis, *The Extended Corporate Mind: When Corporations Use AI to Break the Law*, 98 N.C. L. REV. 893 (2020) [hereinafter Diamantis, *Extended Corporate Mind*]; Matthew U. Scherer, *Of Wild Beasts and Digital Analogues: The Legal Status of Autonomous Systems*, 19 NEV. L.J. 259 (2018); Gabriel Weil, Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence (last updated

law,[39] discussions about appropriate analogies for AI agents in existing law loomed large.[40] Some participants expressed optimism that liability frameworks adapted from either respondeat superior[41] or the closely related law of corporate liability[42] could provide adequate incentives to private actors deploying AI agents. Participants also noted

June 6, 2024) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.4694006; Gabriel Weil, Instrument Choice in AI Governance: Liability as the Indispensable Core (last updated July 8, 2025) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.5283275; Daniel Schwarcz & Josephine Wolff, *The Limits of Regulating AI Safety Through Liability and Insurance: Lessons From Cybersecurity* (Minn. Legal Stud. Res. Paper No. 2025-46, 2025), https://dx.doi.org/10.2139/ssrn.5411062; David C. Vladeck, Essay, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014); Anat Lior, *Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines*, U. CHIC. L. REV. ONLINE (2024), https://lawreview.uchicago.edu/online-archive/holding-ai-accountable-addressing-ai-related-harms-through-existing-tort-doctrines [https://perma.cc/2NSQ-F4EK]; Matthew van der Merwe et al., *Tort Law and Frontier AI Governance*, LAWFARE (May 24, 2024, at 13:38 ET), https://www.lawfaremedia.org/article/tort-law-and-frontier-ai-governance [https://perma.cc/2Y2N-KGZ9]; KETAN RAMAKRISHNAN ET AL., U.S. TORT LIABILITY FOR LARGE-SCALE ARTIFICIAL INTELLIGENCE DAMAGES (2024), https://www.rand.org/pubs/research_reports/RRA3084-1.html; Paulius Čerka et al., *Liability for Damages Caused by Artificial Intelligence*, 31 COMPUT. L. & SEC. REV. 376 (2015); Paulo Padovan et al., *Black Is the New Orange: How to Determine AI Liability*, 31 A.I. & L. 133 (2023); Yonathan A. Arbel et al., *Systemic Regulation of Artificial Intelligence*, 56 ARIZ. ST. L.J. 545 (2024); Mirit Eyal & Yonathan A. Arbel, Racing to Safety: Tax Policy for AI Safety-by-Design (last updated July 30, 2025) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.5181207.

[39] *See, e.g.*, Rebecca Crootof & B.J. Ard, *Structuring Techlaw*, 34 HARV. J.L. & TECH. 347, 387–99 (2021); Matthijs Maas, *AI Is Like… a Literature Review of AI Metaphors and Why They Matter for Policy* (Inst. for L. & A.I., AI Foundations Report 2, 2023), https://law-ai.org/ai-policy-metaphors/ [https://perma.cc/USA7-V48C].

[40] *See, e.g.*, INSTITUTE FOR LAW & AI, *Perspectives on liability | Workshop on Law-Following AI 2025*, at 6:08 (YouTube, Jan. 31, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=KflbJaOa9zw [hereinafter Lior & Schwarcz Workshop Session] (remarks of Anat Lior); Anat Lior, Perspectives on AI Liability (& AI Liability Insurance) 4–5 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI).

[41] *See, e.g.*, Lior & Schwarcz Workshop Session, *supra* note 40, at 8:57–11:17 (remarks of Anat Lior); Lior, *supra* note 40, at 6–7. *See generally* Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043 (2020).

[42] *See, e.g.*, INSTITUTE FOR LAW & AI, *Lessons for LFAI from corporate law | Workshop on Law-Following AI 2025*, at 14:44–15:10 (YouTube, Jan. 31, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=1F7wMygUzow [hereinafter Weitzel & Diamantis Workshop Session] (remarks of Milhailis Diamantis); Diamantis, *Extended Corporate Mind*, *supra* note 38, at 900, 916–31; Diamantis, *Vicarious Liability*, *supra* note 38, at 318–20, 333–34.

that even absent regulation, AI systems and the corporations deploying them will be constrained by existing market and legal constraints.[43]

Others were more skeptical that ex post mechanisms were up to the task. One major source of skepticism starts from the observation that *insurance* will play a large role in determining how the legal rules established in AI liability regimes actually affect the behavior of AI developers and deployers.[44] Since AI developers and deployers are likely to be (and indeed, already are) insured against third-party claims, liability can promote safer behavior only if insurers are able to either price AI-related risks accurately or actively reduce those risks through measures such as offering consulting and risk-management services.[45] However, such "regulation by insurance" in cybersecurity—an arguably structurally similar domain to AI—seems not to have effectively incentivized improved security practices.[46] Although AI-specific insurance policies are beginning to emerge in the market, it remains to be seen whether the insurers offering these policies will be able to gather data that allows them to better design incentives for their policyholders.[47]

Of course, this debate cannot be resolved here. However, we note that, despite these widely varying perspectives on the efficacy of liability for harms caused by AI agents controlled by private actors, few participants contested the premise that liability will be a weaker deterrent for *governmental* AI agents than for governmental human agents. Accordingly, even if new or existing tort doctrines can properly address many risks from private actors, the case for LFAI in the public sector stands on firmer ground.[48]

---

[43] *See* Weitzel & Diamantis Workshop Session, *supra* note 42, at 21:01–21:59 (remarks of Paul Weitzel); Paul Weitzel, Market-Following AI 7–8 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI).

[44] *See* Lior & Schwarcz Workshop Session, *supra* note 40, at 14:47–15:28, 16:23–17:48, 20:31–20:51, 24:52–26:42 (remarks of Daniel Schwarcz); Schwarcz & Wolff, *supra* note 38, at 14–20. *See generally* Daniel Schwarcz, The Limits of Regulating AI Through Liability and Insurance: Lessons from Cyber Security (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI); Kenneth S. Abraham & Catherine M. Sharkey, *The Glaring Gap in Tort Theory*, 133 YALE L.J. 2165 (2024).

[45] *See* Schwarcz & Wolff, *supra* note 38, at 17–19.

[46] *See id.* at 22–38. For a more optimistic perspective on the role of insurance, see Anat Lior, *Insuring AI: The Role of Insurance in Artificial Intelligence Regulation*, 35 HARV. J.L. & TECH. 467 (2022).

[47] *See generally* Anat Lior, *E/Insuring the AI Age: Empirical Insights into Artificial Intelligence Liability Policies*, 31 CONN. INS. L.J. 99 (2025).

[48] *See* O'Keefe et al., *supra* note 1, at 64 ("[W]e 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. We are unsure when LFAI requirements are justified in other domains.") (footnote omitted); *cf.* Weil, *supra* note 38, at 8 ("[Liability] is a weak tool for influencing

The Nuances of Automated Legal Reasoning

For LFAI to work, AI agents must be able to perform reasonably reliable automated legal reasoning:[49]

> [A]n LFAI must be able to determine whether it is obligated to refuse a command from its 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. 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.[50]

Relatedly, the LFAI policy proposal requires some method for determining whether a given AI agent is sufficiently law following.[51] Thus, a significant portion of the workshop was dedicated to discussion of *automated legal reasoning*,[52] providing more thorough coverage of the topic than is available in the LFAI article.[53] As an understanding of the rationale for an AI agent decision is critical for the attribution of liability and the ability to contest such a decision, mechanisms for discerning and recording the reasoning processes of AI agents were also considered.

---

the behavior of national governments . . . ."). A related worry about LFAI in the private sector is that LFAI, as a form of ex ante regulation, would tend to advantage incumbent firms, which could concentrate power. *See* 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]; *cf.* Weitzel & Diamantis Workshop Session, *supra* note 42, at 20:35–21:01, 23:01–23:09 (remarks of Paul Weitzel); Weitzel, *supra* note 42, at 6, 10. This concern would not apply as forcefully to requiring LFAI in governmental settings.

[49] *See generally* O'Keefe et al., *supra* note 1, at 78–80.

[50] *Id.* at 78–79 (footnote omitted).

[51] *See id.* at 119–20; INSTITUTE FOR LAW & AI, *Evaluating AI for legal tasks | Workshop on Law-Following AI 2025*, at 1:39–2:56 (YouTube, Jan. 31, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=6511f8sgC7g [hereinafter Linna Workshop Session]; Daniel W. Linna Jr., Law-Following AI: Evaluation of Artificial Intelligence for Legal Tasks 2–3 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI); INSTITUTE FOR LAW & AI, *Computer science perspectives on AI legal reasoning | Workshop on Law-Following AI 2025*, at 23:05–36:15 (YouTube, Jan. 31, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=1GsTbW-wONgM [hereinafter Atkinson & Guha Workshop Session] (remarks of Neel Guha).

[52] Atkinson & Guha Workshop Session, *supra* note 51; Linna Workshop Session, *supra* note 51.

[53] *See* O'Keefe et al., *supra* note 1, at 78–80.

This area is particularly fast-moving. GPT-5 performs almost twenty percentage points better than GPT-3.5 on Legal Bench,[54] a leading legal AI[55] benchmark. Whereas GPT-3.5 earned no better than a B on law school exams in 2022,[56] OpenAI's o3 earned several A+s this spring.[57] Accordingly, future developments in legal AI (and better understanding of current models' capabilities) could render much of the below discussion dated in the near future.

The Fundamentals of AI Evaluation

A key concept in this area is *evaluation*: "the science of measuring AI behaviors, impacts, and performance."[58] Evaluations might aim to assess the performance of (1) an AI model on its own, (2) an AI model when integrated into some larger AI system, or (3) a human with access to some AI model or system.[59] Evaluations might assess performance relative to some *benchmark*: some normatively desirable level of performance, such as objective correctness or human performance.[60]

---

[54] *LegalBench*, VALS AI (Nov. 13, 2025), https://www.vals.ai/benchmarks/legal_bench-08-26-2025. For background on LegalBench, see Neel Guha et al., *LEGALBENCH: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models*, NIPS '23: PROCS. 37TH INT'L CONF. ON NEURAL INFO. PROCESSING SYS. 44123 (2023), https://dl.acm.org/doi/10.5555/3666122.3668037.

[55] We use "legal AI" as a shorthand for AI systems that perform legal tasks.

[56] *See* Andrew Blair-Stanek et al., GPT-4's Law School Grades: Con Law C, Crim C-, Law & Econ C, Partnership Tax B, Property B-, Tax B, at 2 (May 9, 2023) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.4443471 (citing Jonathan H. Choi et al., *ChatGPT Goes to Law School*, 71 J. LEGAL EDUC. 387, 391 (2022)).

[57] *See* Andrew Blair-Stanek et al., AI Gets Its First Law School A+s (U. Md. Legal Stud. Res. Paper forthcoming) (manuscript at 1), https://dx.doi.org/10.2139/ssrn.5274547.

[58] Atkinson & Guha Workshop Session, *supra* note 51, at 23:05 (remarks of Neel Guha); Neel Guha, Perspective from CS: Legal AI Evaluation 2, (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI).

[59] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 26:59–28:03 (remarks of Neel Guha); *cf.* Linna Workshop Session, *supra* note 51, at 19:13–19:20. For studies on AI-assisted legal work, see, for example, Daniel Schwarcz et al., *AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice* (Minn. Legal Stud. Res. Paper No. 25-16, 2025), https://dx.doi.org/10.2139/ssrn.5162111; Jonathan H. Choi, Amy B. Monahan, & Daniel Schwarcz, *Lawyering in the Age of Artificial Intelligence*, 109 MINN. L. REV. 147 (2024).

[60] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 31:15–32:22 (remarks of Neel Guha).

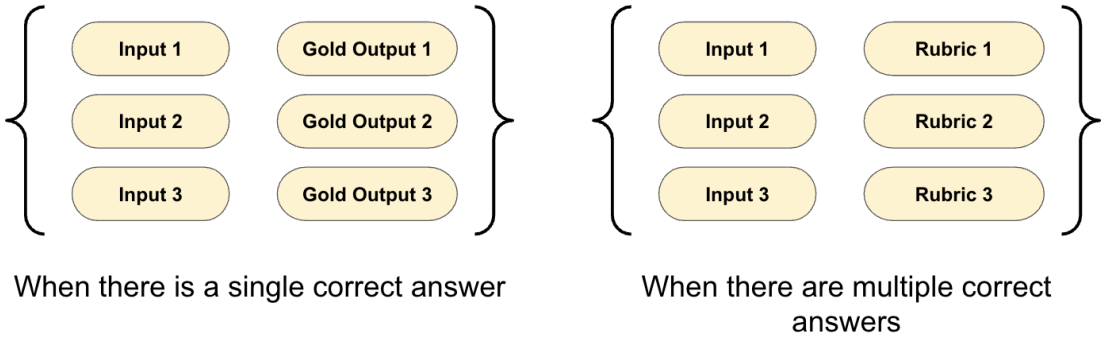

Figure 1: An illustration of AI benchmarks[61]

Not all benchmarks are created equal. When assessing an AI benchmark generally, it is important to assess: (1) whether the inputs are representative of real-world distributions, (2) how subjective the normative outputs are, (3) the temporal stability of the benchmark, (4) the risk that the benchmark "leaks" and contaminates future training data, and (5) the risk that the benchmark is otherwise "gameable."[62]

These factors can both contextualize what, exactly, the benchmark is measuring and help researchers translate benchmark performance into a prediction of real-world performance and impacts.

A major frontier in legal AI evaluation is developing methods of evaluation more nuanced than hallucination rates, particularly for non-objective AI outputs.[63] For example, legal AI systems can be evaluated for, inter alia: (1) *accuracy* on a benchmark; (2)*robustness* "against adversarial attacks and perturbations";[64] (3) *factuality*: whether the output "originates from a verifiable and citable source"; (4) *comprehensiveness*: whether the output "coherently and concisely addresses all aspects of the task"; (5) the *fairness*

[61] Guha, *supra* note 58, at 18.

[62] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 35:22–37:21 (remarks of Neel Guha); Guha, *supra* note 58, at 19–23.

[63] *See* Linna Workshop Session, *supra* note 51, at 14:22–17:11; *see also* Linna, *supra* note 51, at 40–44 (defining hallucination and providing examples). For existing work on hallucination rates in legal AI, see, for example, Varun Magesh et al., *Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools*, 22 J. EMPIRICAL LEGAL STUD. 216 (2025).

[64] On perturbations, see Yu Fan et al., LEXam: Benchmarking Legal Reasoning on 340 Law Exams 23–25 (last updated Oct. 23, 2025) (unpublished manuscript), https://doi.org/10.48550/arXiv.2505.12864; Yiran Hu et al., *J&H: Evaluating the Robustness of Large Language Models Under Knowledge-Injection Attacks in Legal Domain*, 39 PROCS. AAAI CONF. ON A.I. 28106 (2025).

of outputs; (6) the *understandability* of the outputs; and (7) the *transparency* of how the system was developed and how it produced the outputs.[65]

Evaluations of legal AI systems face significant challenges. Since law and legal practice vary by jurisdiction, an inherent limitation of many legal AI evaluations is jurisdiction-dependence.[66] However, there may be much deeper problems. One is data quality. It is sometimes assumed that law is a promising domain for the creation of expert AI systems because of the widespread availability of legal texts (e.g., statutes, case law, court documents, contracts, and legal scholarship). However, there are significant data quality issues for many of these texts.[67] Even legal briefs from the largest law firms frequently contain errors.[68] Recent scholarship has also noted important limitations on the use of judicial opinions as data, since such opinions often intentionally omit the parts of the reasoning process that produced an opinion.[69] Another key limitation is that measuring the relative quality of many of the most important legal outputs, such as briefs, memos, or contracts, is an inherently subjective endeavor notwithstanding high-level objective criteria for quality.

The Role of Explanation in Legal AI

When discussing the potential promise of legal AI systems, care must be taken to identify which legal tasks, exactly, the AI system will be performing. For LFAI, one crucial task is *judgment prediction*:[70] we may wish for LFAIs to base their judgment of whether a contemplated action is legal on a *prediction* of what some court(s) will do.[71]

---

[65] *See* Linna, *supra* note 51, at 50, 57.

[66] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 33:27–33:55 (remarks of Neel Guha).

[67] *See* Linna Workshop Session, *supra* note 51, at 5:12–5:28; Linna, *supra* note 51, at 11–12.

[68] *See* Linna Workshop Session, *supra* note 51, at 5:55–6:06; Linna, *supra* note 51, at 12 (citing Itai Gurari, *Judging Lawyers: Objectively Evaluating Big Law Litigation Departments*, JUDICATA (Jan. 16, 2018), https://blog.judicata.com/judging-lawyers-objectively-evaluating-big-law-litigation-departments-dee7084d86ab).

[69] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 18:39–19:25 (remarks of Katie Atkinson); Katie Atkinson, Computer Science Perspectives on AI Legal Reasoning 17 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI) (citing Courtney M. Cox, *Non-Herculean Data: A Philosophical Intervention in a Technical Debate about Judicial Opinions as Data Sources*, *in* PROCS. 20TH INT'L CONF. ON A.I. & L. (forthcoming 2025) (manuscript at 1–2), https://dx.doi.org/10.2139/ssrn.5471069); Cox, *supra* note 69 (citing Courtney M. Cox, *Super-Dicta*, 173 U. PENN. L. REV. 1575 (2025)).

[70] For an overview of the field of AI judgment prediction, see Masha Medvedeva et al., *Rethinking the Field of Automatic Prediction of Court Decisions*, 31 A.I. & L. 195 (2023).

[71] *See* O'Keefe et al., *supra* note 1, at 125–26.

Of course, while judgment prediction may be useful, it is not the whole picture:[72] *explanation* of legal conclusions plays a crucial role in the legal system.[73] Thus, legal AI systems that provide some form of explanation are also crucial in many contexts.[74] One method of providing such explanations has been the use of argument graphs in which legal arguments (and the relationships between them and legal conclusions) are represented symbolically, enabling logical analysis.[75] Proponents of such techniques argue that it enables provision of "a step-by-step justification" of the overall structure of a legal argument, which can be further supplemented by sources of law or normative rationales justifying each argumentative step.[76] This technique has been used to analyze US trade secret law,[77] the ownership rules for wild animals,[78] the automobile exception

---

[72] *See generally* 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-theorylexicon-the-bad-man-thought-experiment.html [https://perma.cc/4NLC-UFQ8].

[73] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 7:08–7:37 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 6 (citing Henry Prakken, *What Information Does an Algorithmic Legal Judgment Prediction Give?* 6–7 (Dec. 12, 2024), https://webspace.science.uu.nl/~prakk101/talks/JURIX24.pdf [https://perma.cc/8YNH-DCX6]).

[74] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 7:41–8:18 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 7; Linna Workshop Session, *supra* note 51, at 9:05–9:14, 13:00–13:35 (discussing the importance of explanation for improving user comprehension and access to justice). *See generally* Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's Right to an Explanation Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L. REV. 145 (2019); Joshua Gacutan & Niloufer Selvadurai, *A Statutory Right to Explanation for Decisions Generated Using Artificial Intelligence*, 28 INT'L J. L. & INFO. TECH. 193 (2020); Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J. L. &TECH. 84 (2018).

[75] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 8:00–10:03 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 7–8 (citing Sanjay Modgil & Henry Prakken, *The ASPIC+ Framework for Structured Argumentation: A Tutorial*, 5 ARGUMENT & COMPUTATION 31 (2014); Phan Minh Dung, *On the Acceptability of Arguments and Its Fundamental Role in Nonmonotonic Reasoning, Logic Programming and N-Person Games*, 77 A.I. 321 (1995); Latifa Al-Abdulkarim et al., *A Methodology for Designing Systems to Reason with Legal Cases Using Abstract Dialectical Frameworks*, 24 A.I. & L. 1 (2016)); *see also* Gerard A.W. Vreeswijk, *Abstract Argumentation Systems*, 90 A.I. 225 (1997); INSTITUTE FOR LAW & AI, *Use of AI for automated enforcement | Workshop on Law-Following AI 2025*, at 7:21–7:39 (YouTube, Jan. 31, 2026, recorded Aug. 8, 2025), https://www.youtube.com/watch?v=_CCpCoQa4YU [hereinafter Merane Workshop Session] (describing formalization of legal rules into rule-based decision trees); Jakob Merane, Use of AI for Automated Enforcement: What Lessons to Learn for LFAI 9 (Aug. 8, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI).

[76] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 14:10–14:41 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 13.

[77] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 10:03–10:13, 10:28–10:42 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 8–9 (citing Al-Abdulkarim et al., supra note 75, at 13–24).

[78] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 10:13–10:28, 10:42–10:59 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 8–9 (citing Al-Abdulkarim et al., supra note 75, at 24–27).

to the Fourteenth Amendment,[79] and Article 6 of the European Convention on Human Rights.[80] Efforts to combine these symbolic approaches with neural systems like LLMs—and thus reap the benefits of both—are ongoing.[81]

Risks in Automated Compliance

LFAI might be seen as a form of *automated compliance*: LFAIs comply with applicable laws automatically, without the need for ex post enforcement. Due to the opaque and ubiquitous nature of AI systems, it can be costly, even unfeasible, for a regulator to monitor compliance. In such a context, embedding compliance protocols *within* AI systems can form an efficient mechanism to advance lawful AI agent behavior.[82] Another major theme of the workshop explored the possible downsides of this approach to law.

Automated compliance is similar to *perfect enforcement* of the law.[83] But people often chafe at perfect enforcement even of laws they approve of, such as traffic laws.[84] Scholars have noted that the increasing automation of monitoring and enforcement could fundamentally change how laws operate in practice.[85] In the field of privacy and data protection, for example, recent work examined the effects if the internet shifted from selective enforcement to comprehensive automated monitoring of GDPR

---

[79] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 11:00–11:22 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 8–9 (citing Al-Abdulkarim et al., supra note 75, at 27–38).

[80] *See* Atkinson & Guha Workshop Session, *supra* note 51, at 11:38–12:49 (remarks of Katie Atkinson); Atkinson, *supra* note 69, at 13 (citing Joe Collenette et al., *Explainable AI Tools for Legal Reasoning About Cases: A Study on the European Court of Human Rights*, 317 A.I. 103861 (2023)).

[81] *See* Merane Workshop Session, *supra* note 75, at 8:05–8:33; Merane, *supra* note 75, at 14 (citing Karel Kubicek et al., *Automating Website Registration for Studying GDPR Compliance*, PROCS. ACM WEB CONF. 2024 1295, https://doi.org/10.1145/3589334.3645709); *see also* Morgan Gray et al., *Generating Case-Based Legal Arguments with LLMs*, CSLAW '25: PROCS. 2025 SYMP. ON COMPUT. SCI. & L. 160.

[82] For emerging scholarship in this area, see, for example Vinay Kulkarni, et al., *Toward Automated Regulatory Compliance*, 9 CSI TRANSACTIONS ON ICT 95 (2021); Niloufer Selvadurai, *Advancing Lawful AI through Compliance by Design*, 31 COMPUT. & TELECOMMS. L. REV. 35 (2025).

[83] *Cf.* O'Keefe et al., *supra* note 1, at 127 nn.442–43 (relating compliance and enforcement).

[84] *See* Merane Workshop Session, *supra* note 75, at 1:53–2:12; Merane, *supra* note 75, at 3–5 (citing *AI Cameras Catch 590 People Without Seatbelts in Devon and Cornwall*, BBC (Nov. 20, 2022), https://www.bbc.com/news/uk-england-cornwall-63682810 [https://perma.cc/37HU-ETTX]; Jan de Boer, *Zurich Scraps Langstrasse Speed Camera After Issuing Millions in Fines*, I AM EXPAT (Oct. 16, 2024), https://www.iamexpat.ch/expat-info/swiss-news/zurich-scraps-langstrasse-speed-camera-after-issuing-millions-fines [https://perma.cc/X7NX-85QN]).

[85] *See* Woodrow Hartzog et al., *Inefficiently Automated Law Enforcement*, 2015 MICH. ST. L. REV. 1763, 1766–67.

compliance.[86] Such a shift could make rarely enforced rules bite, even where the legislators never intended them to be applied so aggressively.

LFAI policies must therefore contemplate how rigorously LFAIs must obey the law.[87] Per the perfect enforcement literature, the answer should probably not be "perfectly," at least for many laws.[88] Yet even beyond the question of rigor, LFAIs raise deeper challenges. Should an LFAI be a textualist or a purposivist? And when legal rules embody distributive choices (for instance between consumers and corporations) how should those trade-offs be resolved? Law-following requires explicit design decisions about legal interpretation and values,[89] and implicitly requires us to be comfortable with the automatic, large-scale execution of such choices.

Another concern might be that LFAIs could run afoul of antidiscrimination law if not designed carefully.[90] Suppose that an LFAI, through its training process, learns (without being instructed) to use a proxy metric for some protected classification to decide whether to refuse an order.[91] For example, perhaps the model has learned that men are more likely to commit violent crimes than women.[92] In some cases, the LFAI might deny a request from a man that it would comply with if requested by a woman.

---

[86] *E.g.*, Jakob Merane & Alexander Stremitzer, *Automated Private Enforcement: Evidence from the Google Fonts Case* 29–30 (Ctr. for L. & Econ. Working Paper Series No. 4/2025, 2025); Jakob Merane, AI & Law: Automated Private Enforcement of Small Claims 99 (2024) (Ph.D. thesis, ETH Zurich), https://doi.org/10.3929/ethz-b-000725605; Karel Kubíček et al., *Checking Websites' GDPR Consent Compliance for Marketing Emails*, 2022 PROCS. ON PRIV. ENHANCING TECHS. 282, 296.

[87] *See* O'Keefe et al., *supra* note 1, at 127.

[88] *See, e.g.*, 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).

[89] *See* Merane Workshop Session, *supra* note 75, at 15:07–16:46; Merane, *supra* note 75, at 21–22.

[90] INSTITUTE FOR LAW & AI, *Algorithmic resignation | Workshop on Law-Following AI 2025*, at 23:20–23:49 (YouTube, Jan. 31, 2026, recorded Aug. 8, 2025), https://www.youtube.com/watch?v=D5m94opZ7bI [hereinafter Sargeant Workshop Session] (remarks of Holli Sargeant); Holli Sargeant, Algorithmic Resignation 13–14 (Aug. 8, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI); *see also* Holli Sargeant et al., *Unequal Uncertainty: Rethinking Algorithmic Interventions for Mitigating Discrimination from AI* (Aug. 11, 2025) (unpublished manuscript), https://doi.org/10.48550/arXiv.2508.07872.

[91] The term "algorithmic resignation" has been used to refer to this type of mechanism, in which an organization "limit[s] algorithmic assistance in favor of (unaided) human decision-making" based on factors such as uncertainty and user expertise. Bhatt & Sargeant, *supra* note 1, at 100; Sargeant et al., *supra* note 90, at 3.

[92] *E.g.*, ALEXIA COOPER & ERICA L. SMITH, BUREAU JUST. STAT., NCJ 236018, HOMICIDE TRENDS IN THE UNITED STATES, 1980–2008, ANNUAL RATES FOR 2009 AND 2010, 3 (Nov. 2011), https://bjs.ojp.gov/content/pub/pdf/htus8008.pdf.

This could raise antidiscrimination issues,[93] especially in regimes that use a "disparate impact"[94] or "indirect discrimination"[95] standard.

Standard of Care for AI Agents

The workshop featured a rich discussion of the proper standard of care for AI agents, focusing especially on the reasonableness standard.[96] The study of AI reasonableness is attractive for LFAI as a project because reasonableness standards are pervasive in the law;[97] creation of AI agents that behave reasonably, in the legal sense, would thus represent significant progress towards LFAI. The reasonableness standard also recognizes a pluralism of goods anchored in the existing human values, thus making it both a normatively attractive target for study and an exciting technical challenge, especially when compared to the unidimensional measures of performance often used in machine learning.[98]

We can start with the simple question of how well LLMs can already match the reasonableness judgments of humans[99]—a strategy called "silicon sampling" in social science.[100] While LLM outputs are known to imperfectly reflect the views of most populations,[101] the silicon sampling literature nevertheless finds that they *can* produce views

---

[93] *Cf.* Sargeant Workshop Session, *supra* note 90, at 23:20–28:16 (remarks of Holli Sargeant) (describing the problem and other examples); Sargeant, *supra* note 90, at 14.

[94] *See generally* APRIL J. ANDERSON, CONG. RSCH. SERV., IF13057, WHAT IS DISPARATE-IMPACT DISCRIMINATION? (2025).

[95] *See generally Indirect Discrimination*, EMN ASYLUM & MIGRATION GLOSSARY, https://home-affairs.ec.europa.eu/networks/european-migration-network-emn/emn-asylum-and-migration-glossary_en (search in search bar for "indirect discrimination") (last visited Nov. 17, 2025).

[96] Karni Chagal-Feferkorn, Presentation at the 2025 Workshop on Law-Following AI: The Reasonable Algorithm (Aug. 8, 2025) [hereinafter Chagal-Feferkorn Workshop Session]; Karni Chagal-Feferkorn, The Reasonable Algorithm (Aug. 8, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI).

[97] *See* Sunayana Rane, *Position: The Reasonable Person Standard for AI*, ICML'24: PROCS. 41ST INT'L CONF. ON MACHINE LEARNING 42127, 42127, 42129–34 (2024).

[98] *See id.* at 9; *see also* Chagal-Feferkorn Workshop Session, *supra* note 96.

[99] *See generally* Yonathan A. Arbel, The Silicon Reasonable Person: Can AI Predict How People Judge Reasonableness? (U. Ala. Legal Stud. Res. Paper No. 5377475, 2025), https://dx.doi.org/10.2139/ssrn.5377475.

[100] *See id.* at 16; Marko Sarstedt et al., *Using Large Language Models to Generate Silicon Samples in Consumer and Marketing Research: Challenges, Opportunities, and Guidelines*, 41 PSYCH. & MKTG. 1254, 1254–55 (2024).

[101] *See* O'Keefe et al., *supra* note 1, at 112–13.

that correlate remarkably well with human populations.[102] A recent study[103] comparing LLM and human perceptions of reasonableness found remarkable similarities. For example, when evaluating whether failure to take some precaution was negligent, human jurors tend to weigh social factors (that is, how common that precaution is) much more heavily than economic factors (that is, how expensive the precaution would be and how effective it would be at preventing loss), contrary to the general scholarly consensus favoring the latter.[104] Remarkably, LLMs share this bias.[105] Other replications find many other cases in which LLMs mirror human judgments, but also cases in which they differ, sometimes depending on the model used.[106] This line of research suggests that comparing AI agents' and humans' reasonableness judgments is already informative. These comparisons, in turn, might eventually form the basis for concluding that certain AI systems can be trusted to make legally relevant reasonableness judgments as well as humans can.

Yet there are also substantial reasons to doubt that the human baseline is the correct one.[107] "Assessing algorithms by reference to how reasonable people behave [may] set too low of a bar—AI can and should outperform humans on many tasks."[108] This has led to suggestions for the development of a standard of care that incorporates AI agents' unique (and possibly superhuman) competencies. One proposal would hold that an AI behaved unreasonably if it "causes injury more frequently . . . than the combined incident rate for all actors—both human and AI—engaged in the same type of conduct."[109] Another proposal would interrogate both the reasonableness of the AI

---

[102] *See* Arbel, *supra* note 99, at 16–19.

[103] *See id.*

[104] *See* Christopher Brett Jaeger, *The Empirical Reasonable Person*, 72 ALA. L. REV. 887, 904–06 (2021).

[105] *See* Arbel, *supra* note 99, at 22–26.

[106] *See id.* at 26–32.

[107] *See* INSTITUTE FOR LAW & AI, *Responses to Law-Following AI | Workshop on Law-Following AI 2025*, at 2:26–5:18 (YouTube, Jan. 31, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=YaDYY-3K9_s [hereinafter Burri Workshop Session] (remarks of Thomas Burri); Thomas Burri, Responses to Law-Following AI 4 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI); *see also* Diamantis, *Reasonable AI*, *supra* note 38, at 573, 581, 593–602.

[108] Diamantis, *Reasonable AI*, *supra* note 38, at 573; *see also* Karni A. Chagal-Feferkorn, *The Reasonable Algorithm*, 1 U. ILL. J. L. TECH. & POL'Y 111, 145 (2018) ("[J]ust as professionals are held to a higher degree of reasonableness than laypersons, and just as professionals with an established expertise are judged by standards higher than those applied to other professionals, the standard of reasonableness for algorithms might be higher than that for a reasonable 'person' or a 'reasonable professional.'" (footnote omitted)).

[109] Diamantis, *Reasonable AI*, *supra* note 38, at 573.

itself and the reasonableness of the AI developer: the more reasonably the AI itself behaved, the lower the standard of care imposed on its developer.[110]

Interplay Between AI Agents and Principals

"Two-pronged" tort standards that consider the reasonableness of both the AI itself and its developer point to a broader theme: the interplay between AI agents and their principals. Indeed, one way of motivating LFAI and related proposals is to ask: what is the optimal allocation of legal responsibility among the various actors in the chain of causation (e.g., AI developers, deployers, users, and AI itself)?

Another discussion within this theme is the extent to which *fiduciary* principles can accomplish much of what LFAI aims to achieve.[111] Participants overwhelmingly agreed that AI chatbots satisfied each of the elements traditionally used to justify imposition of fiduciary duties.[112] This suggests a strong pro tanto case for treating the developers and providers of AI systems as fiduciaries of their users[113]—or treating AI systems as fiduciaries themselves.[114]

Of course, since fiduciary law is law, we might consider fiduciary AIs as one type of law-following AI. Indeed, fiduciary duties are sometimes understood as entailing an obligation to comply with positive law.[115] But one might extend the fiduciary AI concept further and argue that AI systems with a fiduciary duty to a sufficiently broad set of stakeholders (possibly including broadly conceived stakeholders like "the general public" or even "the Constitution") might, if charged with balancing those duties, achieve the goals of LFAI more effectively, while also reflecting a richer and more

---

[110] *See* Karni A. Chagal-Feferkorn, *How Can I Tell if My Algorithm Was Reasonable?*, 27 MICH. TECH. L. REV. 213, 256–61 (2021).

[111] *See* INSTITUTE FOR LAW & AI, *Fiduciary duties for AI systems | Workshop on Law-Following AI 2025* (YouTube, Jan. 31, 2026, recorded Aug. 8, 2025), https://www.youtube.com/watch?v=VFDs-GOvQx4c [hereinafter Boine & Benthall Workshop Session] (citing Sebastian Benthall & David Shekman, *Designing Fiduciary Artificial Intelligence*, EAAMO '23: PROCS. 3RD ACM CONF. ON EQUITY & ACCESS IN ALGORITHMS, MECHANISMS, & OPTIMIZATION 1 (2023); Claire Boine, *Fiduciary Principles in AI: Utilizing the Duty of Loyalty to Align Artificial Intelligence Systems with Human Goals*, WE ROBOT 2023, https://www.bu.edu/law/files/2023/09/Fiduciary-paper.pdf [http://perma.cc/53BZ-PGGN]).

[112] *See* Boine & Benthall Workshop Session, *supra* note 111, at 1:41–7:22.

[113] *See, e.g.*, Boine, *supra* note 111, at 14–17.

[114] On the latter, see Anthony Aguirre et al., *AI Loyalty: A New Paradigm for Aligning Stakeholder Interests*, 1 IEEE TRANSACTIONS ON TECH. & SOC'Y 128, 130 (2020).

[115] *See* O'Keefe et al., *supra* note 1, at 97 ("[Corporate d]irectors who intentionally cause a corporation to violate positive law breach their duty of good faith.").

nuanced set of considerations than LFAI would allow.[116] The thought here is that agentic AI systems are in fact enmeshed in a network of principal–agent relationships much more complex than the simple principal–agent relationship that LFAI imagines.[117] According to this argument, a networked approach to understanding an AI system's duties—and the inherent conflicts and tensions therein—is more realistic than the command-and-control approach to legal compliance that LFAI imagines.[118]

It is also worth considering how AI agents' "soft skills" will affect how AI agents are actually deployed. In particular, real-world lawyering is more than just logical reasoning: empathizing with clients is a core skill for many forms of legal work.[119] Accordingly, legal AI systems that use empathetic language with their users are perceived as more helpful and trustworthy.[120] This could have numerous implications for LFAI. For example, designers of LFAIs may wish to engineer LFAIs to empathetically deescalate the situation when the user requests that the AI break the law on their behalf. AI developers already have to steer clear of both "overrefusal" (which can degrade user experience and utility without supplying any safety benefit)[121] and "underrefusal" (which can lead to the AI enabling behavior that is foreseeably destructive to the user or others).[122] How well AI developers navigate this tradeoff for illegal requests may significantly influence whether LFAI is perceived by stakeholders as a legitimate and reasonable constraint.[123]

---

[116] *See* Boine & Benthall Workshop Session, *supra* note 111, at 15:38–17:00, 22:19–28:42 (remarks of Sebastian Benthall).

[117] *See id.*

[118] *See id.*

[119] *See* Linna Workshop Session, *supra* note 51, at 9:24–9:45; Linna, *supra* note 51, at 21 (citing Marc Queudot et al., *Improving Access to Justice with Legal Chatbots*, 3 STATS 356 (2020)).

[120] *See* Linna Workshop Session, *supra* note 51, at 12:34–12:41; Linna, *supra* note 51, at 25 (citing Sabine Brunswicker et al., *The Impact of Empathy in Conversational AI: A Controlled Experiment with a Legal Chatbot*, PROCS. 57TH HAWAII INT'L CONF. ON SYS. SCIS. 455 (2024)).

[121] *See, e.g.*, Mahavir Dabas et al., *Just Enough Shifts: Mitigating Over-Refusal in Aligned Language Models with Targeted Representation Fine-Tuning*, PROCS. 42ND INT'L CONF. ON MACH. LEARNING 11846, 11846 (2025).

[122] *See, e.g.*, Andrew Clark, *The Ability of AI Therapy Bots to Set Limits with Distressed Adolescents: Simulation-Based Comparison Study*, 12 JMIR MENTAL HEALTH e78414 (2025).

[123] *Cf.* Merane Workshop Session, *supra* note 75, at 22:48–23:27 (identifying potential responses to an illegal request); Merane, *supra* note 75, at 28; Bhatt & Sargeant, *supra* note 1, at 102.

Evaluating AI Mental States

Many laws have a mental state as an element. To be able to say that an AI agent violated the law, we likely need some way of evaluating whether it acted with the requisite mental state.[124] To that end, the workshop featured a session on different approaches to thinking about AI intentionality.[125]

Drawing on a distinction in philosophy of mind, a session at the workshop distinguished between *realist* and *interpretivist* approaches to inferring or imputing mental states.[126] Realists try to assess the subject's actual mental state as subjectively experienced by that subject.[127] Interpretivists, on the other hand, conclude that the subject's mental state is that mental state that best serves as a "coherent explanation for future behavior."[128] Interpretivism is often associated with Daniel Dennett's "intentional stance":[129] "the strategy of prediction and explanation that attributes beliefs, desires, and other 'intentional' states to systems—living and nonliving—and predicts future behavior from what it would be rational for an agent to do, given those beliefs and desires."[130]

When polled on their views as to the appropriate approach for humans, workshop participants were evenly split between realism and interpretivism.[131] For AI systems, by contrast, most participants endorsed some form of interpretivism.[132] Views on the latter varied widely, of course, with many participants remaining skeptical of attributing mental states to AI systems.[133]

---

[124] *See* O'Keefe et al., *supra* note 1, at 86–93. However, there is the theoretical possibility of creating laws for AI agents that do not have mental state elements. *See id.* at 92.

[125] *See* INSTITUTE FOR LAW & AI, *AI mental states: Takeaways from intentionality workshop | Workshop on Law-Following AI 2025* (YouTube, Jan. 31, 2026, recorded Aug. 8, 2025), https://www.youtube.com/watch?v=0ukHp98RlcQ [hereinafter Cheong Workshop Session]. For a key piece of research that laid the groundwork for modelling intention in agent systems, see MICHAEL E. BRATMAN, INTENTION, PLANS, AND PRACTICAL REASON (1987).

[126] *See* Cheong Workshop Session, *supra* note 125, at 8:58–9:39. These categories and the discussion in this section necessarily simplify a complicated area of philosophy of mind.

[127] *See id.*

[128] *See id.*

[129] *See id.* at 13:26–15:22.

[130] Daniel C. Dennett, *Précis of* The Intentional Stance, 11 BEHAV. & BRAIN SCIS. 495, 495 (1988); *see also* CHOPRA & WHITE, *supra* note 1, at 12–13, 146 (discussing using the intentional stance to ascribe mental states to AIs).

[131] Cheong Workshop Session, *supra* note 125, at 9:39–10:10.

[132] *See id.* at 10:32–12:16.

[133] *See id.* at 16:52–18:16.

Legal systems around the world are already grappling with AI mental states, and how they relate to the mental states of AI systems' developers and users. The European Commission's Guidelines on Prohibited AI Practices forbids "purposefully manipulative techniques," including "AI systems that manipulate individuals *without any human intending them to do so*."[134] The district court in *Garcia v. Character Technologies, Inc.*,[135] in which a chatbot is alleged to have caused a teen's suicide,[136] had to confront the question of how AI speech related to developers' intentions. The court was asked to hold that the chatbot's output was protected speech.[137] It declined to do so,[138] giving serious weight to the proposition that protected "speech" must come from "a human being with First Amendment rights [who has] made an inherently expressive '*choice*'" to publish that content.[139] We should expect questions of AI mental states to become increasingly pressing for courts and litigants as AI systems become increasingly capable of generating content and taking actions that could give rise to legal actions.

Legal Status for AI Agents

The *Law-Following AI* article proposes that LFAIs be treated as a new type of legal entity: a "legal actor" on which the law imposes duties but not rights.[140] A number of conversations at the workshop explored the implications of this aspect of LFAI. A recurring concern with this proposal was that it was obfuscatory, possibly deflecting attention away from the developers of AI systems.[141]

---

[134] Commission Guidelines on Prohibited Artificial Intelligence Practices Established by Regulation (EU) 2024/1689 (AI Act), C(2025) 5052 final, ¶¶ 58, 69 (Jul. 29, 2025) (emphasis added); *see* Cheong Workshop Session, *supra* note 125, at 24:04–25:38.

[135] Order on Motion to Dismiss, Garcia v. Character Techs., Inc., No. 6:24-cv-01903-ACC-UAM, (M.D. Fla. Filed May 21, 2025), ECF No. 115.

[136] *See id.* at 6–9.

[137] *See id.* at 27–32.

[138] *See id.* at 31.

[139] *Id.* at 31 (quoting Moody v. NetChoice, 603 U.S. 707, 746 (2024) (Barrett, J., concurring)) (emphasis added). To be clear, however, the court was not necessarily holding that the AI lacked the relevant intent. It was merely holding that only content ultimately traceable to the authorial "choice" of some human was protected. *See id.* at 31 (quoting *Moody*, 603 U.S. at 746 (Barrett, J., concurring)).

[140] *See* O'Keefe et al., *supra* note 1, at 83–86.

[141] *E.g.*, Burri Workshop Session, *supra* note 107, at 10:23–12:34; *see also, e.g.*, Daniel Leufer, *Why We Need to Bust Some Myths About AI*, 1 PATTERNS 100124, at 1, 2 (2020) ("[T]he problem is that the ascription of agency to AI *masks the human agency behind certain processes*."); Jack M. Balkin, *The Three Laws of Robotics in the Age of Big Data*, 78 OHIO ST. L.J. 1217, 1224–25 (2017). The *Law-Following AI* article anticipates this criticism and responds as follows:

But there was also skepticism about the legal actor proposal from the other direction, holding that it did not go far enough. One repeated concern was that giving LFAIs legal duties without legal rights of some sort would be unworkable: to fairly adjudicate whether an AI system violated a legal duty, we might need to give that AI agent some procedural (e.g., the right to counsel, the right to appeal) and substantive (e.g., rights that give rise to an affirmative defense) rights.[142] This seems correct. Future work on LFAI should more carefully explore which rights AI agents will need to be given if we are to impose legal duties on them. Nevertheless, it still seems possible to give AI agents many fewer rights than most legal persons.[143]

An even more radical set of proposals argued for giving AI agents a more fulsome set of private-law rights.[144] The thought is that granting AI agents secure property rights could encourage positive-sum trade between humans and AI agents with their own misaligned goals, thus mitigating the risk of conflict between them.[145] Since LFAI relies, in part, on ensuring that AI agents are aligned,[146] these proposals might be seen as an alternative to LFAI, in case LFAI is not timely and effectively implemented.

> To ascribe duties to AI agents is not to deflect moral and legal accountability for their developers and users, as some critics have charged. 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—creating and directing a new type of actor—so as to reach a better conclusion as to the nature of their responsibilities. Our proposition is that those developers and principals should have an obligation to, among other things, ensure that their AI agents are law following.

O'Keefe et al., *supra* note 1, at 85–86 (footnotes omitted).

[142] *See, e.g.*, Weitzel & Diamantis Workshop Session, *supra* note 42, at 6:12–6:41 (remarks of Milhailis Diamantis); Boine & Benthall Workshop Session, *supra* note 111, at 20:59–21:06 (remarks of Claire Boine).

[143] *Cf.* Delgado, *supra* note 1, at 4–5 (identifying types of entities with very limited bundles of rights and duties as possibly analogous for LFAI).

[144] *See* INSTITUTE FOR LAW & AI, *AI systems as holding rights and duties | Workshop on Law-Following AI 2025*, at 2:34–3:38 (YouTube, Jan. 31, 2026, recorded Aug. 8, 2025), https://www.youtube.com/watch?v=XrVTMqefjyA [hereinafter Salib & Arbel Workshop Session] (remarks of Peter Salib); Salib & Goldstein, *supra* note 33, at 4, 7–9, 33–68; Simon Goldstein & Peter Salib, AI Rights for Human Flourishing (last updated Aug. 7, 2025) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.5353214.

[145] *See* Salib & Arbel Workshop Session, *supra* note 144, at 14:29–19:03 (remarks of Peter Salib); *see also* Henry A. Thompson, Some Economics of Artificial Superintelligence (last updated Nov. 15, 2025), http://dx.doi.org/10.2139/ssrn.5728702.

[146] *See* O'Keefe et al., *supra* note 1, at 108–16.

Case Study: Law-Following AI and International Humanitarian Law

LFAI owes a significant intellectual debt to the rich scholarly literature about whether and how AI-enabled systems, especially autonomous weapons, might be made to comply with international humanitarian law (IHL).[147] Autonomous weapons present a rich domain for work on LFAI for obvious reasons. Assessments in identifying proper targets, for example, may reflect the pinnacle of high-stakes legal analysis. Autonomous weapons systems with excessively permissive targeting standards will impose unnecessary harm to civilian persons and objects. On the other hand, an excessively cautious approach to automated IHL compliance could have catastrophic tactical consequences, such as an autonomous weapon system erroneously refusing to engage a lawful target, thereby endangering human allies relying on its functionality.

Many States have already endorsed LFAI-like principles for autonomous weapons, through instruments such as the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, which declares that "use of AI in armed conflict must be in accord with States' obligations under international humanitarian law, including its fundamental principles."[148] Yet, IHL rules and principles are unmistakably difficult to formally represent. For example, IHL's proportionality principle hinges on an assessment of "military advantage" a concept for which it has been asserted that "no abstract calculations [are] possible."[149] Likewise, the scope of behaviors that render civilians "direct participants" in hostilities and thus lawful targets is often considered by

147 *See, e.g.*, PAUL SCHARRE, ARMY OF NONE: AUTONOMOUS WEAPONS AND THE FUTURE OF WAR 282–83 (2018); Ashley Deeks, *Coding the Law of Armed Conflict: First Steps*, *in* THE FUTURE LAW OF ARMED CONFLICT 41 (Matthew C. Waxman & Thomas W. Oakley eds., 2022).

148 Bureau of Arms Control & Nonproliferation, U.S. Dep't of State, Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (Nov. 9, 2023), https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/ [https://perma.cc/9GM3-S79U].

149 INSTITUTE FOR LAW & AI, *Law-Following AI in International Humanitarian Law | Workshop on Law-Following AI 2025*, at 27:48–28:08 (YouTube, Feb. 6, 2026, recorded Aug. 7, 2025), https://www.youtube.com/watch?v=Dy07B2pBGZE [hereinafter Sullivan Workshop Session] (referring to IHL proportionality calculations); Scott Sullivan, Law-Following AI and the Laws of War 25 (Aug. 7, 2025) (unpublished presentation slides) (on file with the Institute for Law & AI) (quoting Int'l Comm. Red Cross, Practice Relating to Rule 14: Proportionality in Attack (Germany), CUSTOMARY IHL DATABASE, https://ihl-databases.icrc.org/en/customary-ihl/v2/rule14?country=de [https://perma.cc/9D9J-45F2]).

States to be "undefined and largely undefinable."[150] These difficulties and tensions are emblematic of the challenges posed by the introduction of LFAs for governmental functions more generally.

One session at the Workshop explored the implications of this literature for LFAI, especially in light of the current wave of autonomous weapons being deployed in ongoing conflicts.[151] One key background observation is that IHL primarily regulates the *use* of weapons, rather than their design and development.[152] For example, the International Court of Justice has held that the legality of nuclear weapons—the most destructive technology ever created by humans—rested primarily on the context of their use rather than the intrinsic destructiveness of their design.[153] There are, of course, exceptions. "Weapons that are, by their nature, indiscriminate" can be proscribed on the basis of their design.[154] An example of this is biological weapons, which by their nature cannot be used consistent with the principle of distinction.[155] Another example is weapons that are designed such that they cause superfluous injury or unnecessary suffering in contravention of the principle of necessity, such as poisoned weapons.[156] Finally, states must assess weapons for legality prior to deployment, accounting for both intended and foreseeable misuse.[157]

These distinctions are relevant to LFAI insofar as LFAI primarily proposes that AI agents be regulated through design rather than their use.[158] While there is ample

---

[150] *See* Scott Sullivan & Iben Ricket, *Targeting in the Black Box*, 16th INTL. CONF. ON CYBER CONFLICT: OVER THE HORIZON 207, 214 (2024) (citing U.S. DEP'T OF THE NAVY, COMDTPUB P5800.7A , THE COMMANDER'S HANDBOOK ON THE LAW OF NAVAL OPERATIONS § 8.2.2 (2022)).

[151] *See generally* Sullivan Workshop Session, *supra* note 149.

[152] *See* Sullivan Workshop Session, *supra* note 149, at 10:59–13:17; Sullivan, *supra* note 149, at 11–14.

[153] *See* Sullivan Workshop Session, *supra* note 149, at 11:25–13:17; Sullivan, *supra* note 149, at 12 (citing Legality of the Threat or Use of Nuclear Weapons, Advisory Opinion, 1996 I.C.J. 226, ¶ 97 (July 8)). *But cf.* Colin B. Picker, *A View from 40,000 Feet: International Law and the Invisible Hand of Technology*, 23 CARDOZO L. REV. 149, 175 (2001) (arguing that the ICJ opinion, properly read, renders many forms of nuclear weapons illegitimate).

[154] Sullivan Workshop Session, *supra* note 149, at 13:25–14:13; Sullivan, *supra* note 149, at 14.

[155] *See* Sullivan Workshop Session, *supra* note 149, at 14:13–14:40; Sullivan, *supra* note 149, at 14.

[156] *See* Sullivan Workshop Session, *supra* note 149, at 14:40–15:23; Sullivan, *supra* note 149, at 14.

[157] *See* Sullivan Workshop Session, *supra* note 149, at 15:43–15:55; Sullivan, *supra* note 149, at 14; *see also* Antonio Coco & Talita Dias, '*Handle with Care': Due Diligence Obligations in the Employment of AI Technologies*, *in* RESEARCH HANDBOOK ON WARFARE AND ARTIFICIAL INTELLIGENCE 234, 245 (Robin Geiß & Henning Lahmann eds., 2024) (arguing that Article 1 common to the Geneva Conventions covers both the design and deployment of autonomous weapons); Aleksi Kajander et al., *Making the Cyber Mercenary—Autonomous Weapons Systems and Common Article 1 of the Geneva Conventions*, *in* 12TH INTERNATIONAL CONFERENCE ON CYBER CONFLICT 79, 84 (Tatiana Jančárková et al. eds., 2020).

[158] *See* O'Keefe et al., *supra* note 1, at 93–108.

precedent for such design-based regulation within IHL, IHL, like most of international law,[159] typically defaults to technology-neutral rules of state conduct. However, there is reason to believe that some of the distinctive technical properties associated with AI—such as model opacity, emergent behavior, vulnerability to counter-AI, and propensity to systematic error—create obstacles to IHL compliance that cannot be effectively neutralized at the point of use. This suggests that LFAI would fit most comfortably in IHL in circumstances where reliance on use regulations cannot sufficiently ensure compliance with IHL when considering a State's envisioned use of the AI systems in question.[160] It seems plausible that, given AI agents' high degree of autonomy and deploying States' limited supervisory capabilities, it may indeed be hard to achieve such assurances with use-based restrictions alone, therefore possibly justifying design-based regulation like LFAI.[161]

Conclusion

As the above discussion reveals, the LFAI research agenda still contains many open questions. Many of these questions relate to the proper scope of LFAI: When can other tools address the harms LFAI aims to prevent? Would LFAI exhibit the same flaws as prior attempts at automated enforcement? Does LFAI objectionably deflect responsibility from the developers of AI systems? Other questions relate to the design of LFAIs and related policies: To what extent is symbolic representation of legal rules necessary in the age of LLMs? How would a "reasonable" AI agent behave? What is the exact bundle of duties and rights we should give to LFAIs?

The workshop did not generate widespread consensus on answers on these questions. But that is to be expected. LFAI, at its most ambitious, contemplates the restructuring of the legal order to account for the introduction of a powerful new type of actor. Such a restructuring is bound to raise thorny questions, many of which will have no obvious answer.

Nevertheless, the overall tenor of discussion at the workshop was optimistic. The questions identified may be difficult, but participants generally felt that the topic was important and generative. Society must address, one way or another, novel (or newly amplified) risks from AI agents. Designing AI agents to respect legal rules will almost

---

[159] *See* MATTHIJS M. MAAS, ARCHITECTURES OF GLOBAL AI GOVERNANCE: FROM TECHNOLOGICAL CHANGE TO HUMAN CHOICE 251–53 (2025).

[160] Sullivan Workshop Session, *supra* note 149, at 18:01–18:48; Sullivan, *supra* note 149, at 17.

[161] *See* Sullivan Workshop Session, *supra* note 149, at 19:43–21:17.

certainly be one method for doing so; indeed, it already is.[162] As long as AI companies and policymakers continue to treat the law as a source of values for AI systems, scholars will need to grapple with the core questions that LFAI raises. This line of inquiry seems likely to, in the fullness of time, yield significant and actionable insights, just as scholarship on the nature and treatment of business entities has enabled market economies to capitalize on their enormous benefits while managing, however imperfectly, some of their most serious risks. The main uncertainty, then, is not whether further research on LFAI will be useful, but rather whether such insights will keep pace with advances in AI technology. It is our hope that, through events like the workshop and documents like this one, we can increase the odds that humanity will develop the ideas we need to safeguard human welfare and the rule of law in this transformative era.

[162] *See* O'Keefe et al., *supra* note 1, at 81–82.

About the Authors

Authors are ordered alphabetically by last name. The four workshop organizers are listed first.

Cullen O'Keefe, Director of Research, Institute for Law and AI; Research Affiliate, Centre for the Governance of AI

Christoph Winter, Assistant Professor of Law and AI, University of Cambridge; Director, Institute for Law & AI

Matthijs Maas, Senior Research Fellow, Institute for Law & AI; Research Affiliate, Leverhulme Centre for the Future of Intelligence, University of Cambridge

Janna Tay, Research Scholar, Institute for Law and AI

Yonathan Arbel, Rose Professor of Law University of Alabama School of Law; Co-Director of the Center for Law & AI Risk

Katie Atkinson, Professor of Computer Science, Associate Pro-Vice-Chancellor, and Director of the Interdisciplinary Centre for Sustainability Research at the University of Liverpool

Sebastian Benthall, Research Director, Information Law Institute, New York University School of Law

Jack Boeglin, Sharswood Fellow, University of Pennsylvania Carey Law School; Research Affiliate, Institute for Law & AI

Claire Boine, Assistant Professor in Digital Governance, School of Transnational Governance, European University Institute

Thomas Burri, Professor of International Law and European Law, Law School, University of St. Gallen

Nicholas Caputo, Legal Researcher, Oxford Martin AI Governance Initiative

Karni Chagal-Feferkorn, Assistant Professor of Instruction, University of South Florida Bellini College of AI, Cybersecurity and Computing

Alan Chan, Research Fellow, Centre for the Governance of AI

Inyoung Cheong, Postdoctoral Researcher, Princeton Center for Information Technology Policy

Mihailis Diamantis, Ben V. Willie Professor in Excellence, University of Iowa College of Law

Talita Dias, Senior Research Fellow & Project Director, University of Exeter

Michael Dorff, Executive Director, Lowell Milken Institute for Business Law and Policy, UCLA School of Law

Kevin Frazier, AI Innovation and Law Fellow & Lecturer, University of Texas at Austin School of Law

Neel Guha, PhD Candidate, Stanford Computer Science

Julius Mentis Hattingh, JSD Candidate, Yale Law School; 2025 Summer Research Fellow (Legal Frontiers), Institute for Law & AI

Maarten Herbosch, Senior Research Scholar, Institute for Law & AI; Assistant Professor in AI and Law, KU Leuven

Apostolos Latsonas, LLM Candidate, Harvard Law School; 2025 Summer Research Fellow (EU Law), Institute for Law & AI

Dan Linna, Senior Lecturer and Director of Law and Technology Initiatives, Northwestern Pritzker School of Law & McCormick School of Engineering

Anat Lior, Assistant Professor of Law, Drexel University; Assistant Professor, Center for Science, Technology, and Society (by courtesy); Drexel University; AI Schmidt Affiliated Scholar, Jackson School for Global Affairs, Yale University; Affiliated Fellow, Information Society Project (ISP), Yale Law School

Jakob Merane, Senior Researcher & Lecturer, ETH Zurich, Center for Law & Economics

Sunayana Rane, Department of Computer Science, Princeton University; University of Chicago Law School

Peter Salib, Assistant Professor of Law, University of Houston Law Center; Executive Co-Director, Center for Law & AI Risk; Senior Research Affiliate, Institute for Law & AI

Holli Sargeant, Research Fellow in Law, St John's College, University of Cambridge

Daniel Schwarcz, Fredrikson & Byron Professor of Law & Distinguished University Teaching Professor, University of Minnesota Law School

Niloufer Selvadurai, Professor of Technology Law, Macquarie University; Director of Policy, Applied AI Research Centre

Henry A. Thompson, Assistant Professor of Economics, University of Mississippi; 2025 Summer Research Fellow (Legal Frontiers), Institute for Law & AI

Paul Weitzel, Schmid Professor for Excellence in Teaching & Assistant Professor of Law, University of Nebraska–Lincoln

Spencer Williams, Professor of Law, California Western School of Law

Additional thanks to several participants in the workshop who contributed greatly to the discussion but were unable to participate in the writing process: Bryan Druzin, Anthony Niblett, and Madalina Nicolai.

About Lawfare

Lawfare is a non-profit independent multimedia publication and research organization dedicated to "Hard National Security Choices." We offer non-partisan, timely analysis of thorny legal and policy issues through our written, audio, and other content—all of which you can find at www.lawfaremedia.org. We aim to improve the discourse on the law and policy of national security with a relentless focus on substantive issues that matter—in a fashion that is useful to policymakers and practitioners, but also accessible to anyone who wants to access it. Lawfare is proud to provide all our content free of charge and without a subscription.

You can support our work by donating at www.patreon.com/lawfare. For press and booking inquiries, email Lawfare at press@lawfaremedia.org.