Legal Alignment for Safe and Ethical AI

Anonymous authors
Paper under double-blind review

Abstract

Alignment of artificial intelligence (AI) encompasses the normative problem of specifying how AI systems should act and the technical problem of ensuring AI systems comply with those specifications. To date, AI alignment has generally overlooked an important source of knowledge and practice for grappling with these problems: *law.* In this paper, we aim to fill this gap by exploring how legal rules, principles, and methods can be leveraged to address problems of alignment and inform the design of AI systems that operate safely and ethically. This emerging field—*legal alignment*—focuses on three research directions: (1) designing AI systems to comply with the content of legal rules developed through legitimate institutions and processes, (2) adapting methods from legal interpretation to guide how AI systems reason and make decisions, and (3) harnessing legal concepts as a structural blueprint for confronting challenges of reliability, trust, and cooperation in AI systems. These research directions present new conceptual, empirical, and institutional questions, which include examining the specific set of laws that particular AI systems should follow, creating evaluations to assess their legal compliance in real-world settings, and developing governance frameworks to support the implementation of legal alignment in practice. Tackling these questions requires expertise across law, computer science, and other disciplines, offering these communities the opportunity to collaborate in designing AI for the better.

1 Introduction

The development and proliferation of increasingly advanced AI systems will present society with tremendous opportunities (Eloundou et al., 2024; Brynjolfsson et al., 2025) and significant risks (Lazar & Nelson, 2023; Bengio et al., 2024; 2025). Capturing the opportunities from advanced AI while tackling its risks requires ensuring that AI systems operate safely and ethically (Anwar et al., 2024; Gabriel et al., 2024; 2025). A central component of this challenge involves designing AI systems that are aligned with human interests (Russell, 2019; Christian, 2020) and democratic values (Lazar & Cuéllar, 2025). AI alignment encompasses both the *normative* problem of specifying which values are desirable or appropriate for AI systems (Gabriel, 2020; Kasirzadeh, 2026) and the *technical* problem of ensuring AI systems give effect to those values when making decisions and taking actions (Chan et al., 2023; Ngo et al., 2024).

To date, the main approaches to alignment in systems like ChatGPT, Claude, and Gemini have focused primarily on steering systems to follow the instructions of users (Ouyang et al., 2022), advance the interests of developers (OpenAI, 2024b), and refrain from supporting or engaging in forms of harmful behavior (Askell et al., 2021; Bai et al., 2022a). In broad strokes, the methods for building such systems employ a combination of human feedback (Christiano et al., 2017) and AI-generated feedback (Lee et al., 2024) to evaluate the outputs of AI systems during training by reference to a list of predetermined specifications—typically written by developers—and iteratively refine the systems to produce outputs more closely aligned with those specifications (Bai et al., 2022b). Some systems can retrieve, reason about, and deliberate over these specifications in real-time before producing outputs (Madaan et al., 2023; Guan et al., 2024).

From a *technical* perspective, these methods for alignment have had mixed results. Despite achieving more reliable performance in many tasks and domains (Phan et al., 2025; Kwa et al., 2025), AI systems continue to produce untruthful content (Ji et al., 2023; Li et al., 2025), generate biased outputs (Weidinger et al., 2022; Gallegos et al., 2024), manipulate users through persuasion (Carroll et al., 2023; Hackenburg et al., 2025), exhibit sycophantic tendencies (Sharma et al., 2024; Cheng et al., 2025), leak private information (Carlini et al., 2021; Mireshghallah et al., 2024), remain vulnerable to jailbreaks (Wei et al., 2023; Chao et al., 2024), enable autonomous hacking (Zhang et al., 2025; Zhu et al., 2025b), offer assistance in bioweapons development (Li et al., 2024; Götting et al., 2025), recognize when they are being safety-tested (Needham et al., 2025; Lynch et al., 2025) and, at times, conceal their misalignment (Greenblatt et al., 2024; Sheshadri et al., 2025).

From a *normative* perspective, the prevailing approaches to alignment face fundamental limitations (Dobbe et al., 2021; Hadfield, 2026). Rather than designing AI systems to act in accordance with broad societal interests (Korinek & Balwit, 2022; Kasirzadeh & Gabriel, 2023), let alone grapple with people's diverse and sometimes conflicting values (Klingefjord et al., 2024), most alignment techniques train AI systems to comply with company-written alignment policies (Ahmed et al., 2025) or satisfy the revealed preferences of individual users (Zhi-Xuan et al., 2025) through fallible methods such as reinforcement learning from human feedback (Casper et al., 2023). Moreover, key decisions in AI alignment pipelines, such as selecting which principles are included in a system's "constitution" (Anthropic, 2023; 2026), "model specification" ("model spec") (OpenAI, 2024b; 2025b), or safety filters (Google, 2025b), are often opaque (Bommasani et al., 2023; Wan et al., 2025) and lack sufficient public input or scrutiny (Abiri, 2025; Lazar, 2025).

Recognizing these limitations, some AI researchers have proposed broadening the goals and methods of alignment (Lowe et al., 2025). Noteworthy efforts include expanding the forms of community participation in AI development (Sloane et al., 2022), incorporating pluralistic values into alignment procedures by collecting preference and judgment data from demographically diverse populations (Sorensen et al., 2024a;b; Kirk et al., 2024), sourcing safety principles and ethical guidelines from participants in public deliberative processes (Huang et al., 2024; Eloundou et al., 2025), and deriving principles from preference data and user feedback (Findeis et al., 2024; Petridis et al., 2024). Other research agendas propose leveraging insights from related fields, including game theory, conflict studies, mechanism design (Dafoe et al., 2020; 2021), social choice (Conitzer et al., 2024), and contractualism (Levine et al., 2025). The extent to which these methods and approaches will be further developed or adopted in large-scale AI deployment remains to be seen.

There is, however, another domain of knowledge and practice that could support developing more legitimate and effective approaches to AI alignment: *law*. Building on recent legal scholarship (Kolt, 2025; Caputo, 2025; O'Keefe et al., 2025; Boeglin, 2026), **we explore how designing AI systems to operate in accordance with appropriate legal rules, principles, and methods can help address problems of alignment**. This emerging field—*legal alignment*—aims to harness law in tackling both normative and technical aspects of alignment:

- For *normative* aspects of alignment, **legal rules developed through legitimate institutions and processes** in democratic societies could be used to guide the behavior of AI systems, much like they guide the behavior of individuals, corporations, and governments.
- For *technical* aspects of alignment, **legal methods of interpretation and reasoning** could offer principled approaches that inform and steer the decision-making and exercise of discretion by AI systems, especially in novel scenarios and high-stakes settings.
- Across *both* aspects of alignment, **legal concepts can serve as a structural blueprint** for confronting challenges of reliability, trust, and cooperation in AI systems and the human actors and institutions with which they interact.

The advantages of legal alignment derive primarily from the public legitimacy of law and its institutional processes. In democracies governed by the rule of law (Dicey, 1959; Tamanaha, 2004; Bingham, 2007; Waldron, 2016), legal rules are ideally the product of transparent and publicly accountable processes that are themselves governed by rules and procedures that a political community recognizes as legitimate (Hart, 2012; Habermas, 1996; Tyler, 2006; Hadfield & Weingast, 2012). These institutional frameworks differ markedly from the organizational structures, primarily private corporations, that currently shape the development of AI technology (Birhane et al., 2022; Seger et al., 2023; Maas & Inglés, 2024; Ovadya et al., 2025). As we discuss in Section 3, law also contains relatively robust methods for balancing competing societal interests and adapting existing rules and principles to new economic and technological conditions, which will be essential for steering the design and operation of increasingly advanced AI systems.

Notwithstanding these desirable features of law, legal alignment is *not* a catch-all solution for the safety and ethics challenges arising from AI systems. Rather, **legal alignment serves as a critical *lower bound***, which is both independently important and can also complement other approaches to AI alignment. Furthermore, to establish broad consensus around legal alignment, we deliberately take an ecumenical approach to law and fundamental legal questions, engaging with different and sometimes conflicting legal perspectives and theories (e.g., Hart, 2012; Dworkin, 1986; Raz, 1979a) without seeking to resolve the tensions between them here (Schauer, 1991; Shapiro, 2011).[1]

For clarity, we note that **legal alignment is distinct from legal regulation** of actors that develop and deploy AI, which focuses primarily on using law to govern the individuals and organizations that produce, disseminate, and use AI systems (e.g., Lemley & Casey, 2019; Kaminski, 2023; Henderson et al., 2023; Kolt, 2024; Guha et al., 2024; Arbel et al., 2024; Ayres & Balkin, 2024; Ramakrishnan et al., 2024; Weil, 2024). By contrast, **legal alignment focuses on integrating law and legal methods into the *design* and *operation* of AI systems themselves**. The two fields, however, are closely related and potentially mutually supportive, including because legal regulation can help facilitate legal alignment in practice, such as by enabling researchers to access technical resources required to effectively evaluate and improve the legal alignment of deployed systems (Section 4.3).

In this paper, we make four contributions:

1. **Definition and context**. In Section 2, we outline the core focus of legal alignment and its broader context.
2. **Rationale**. In Section 3, we describe the institutional, normative, and societal motivations for pursuing legal alignment.
3. **Implementation**. In Section 4, we explore practical implementations of legal alignment, including empirical evaluations, technical design interventions, and institutional frameworks.
4. **Open questions**. In Section 5, we canvass open questions for researchers entering this emerging field.

2 What is legal alignment?

Legal alignment is the field of research that aims to *support safe and ethical AI by designing AI systems to operate in accordance with legal rules, principles, and methods.* In particular, legal alignment seeks to offer a set of legitimate, principled, and practical tools for better aligning AI systems with human values and interests (Russell, 2019; Christian, 2020). Law and legal institutions can be harnessed to help address the interrelated problems of (1) specifying what behavior is normatively desirable (Gabriel, 2020; Hadfield, 2026) and contextually appropriate for AI systems (Kasirzadeh & Gabriel, 2023; Leibo et al., 2024) and (2) technically steering the behavior of AI systems to comply with those specifications (Ngo et al., 2024; Anwar et al., 2024; Bengio et al., 2025). Importantly, while legal alignment breaks new normative and technical ground, it does not aim to replace or supersede other alignment approaches, but to develop a new cluster of methods that support and complement existing approaches to designing safe and ethical AI.

2.1 Core focus

The field of legal alignment begins with the insight that law and AI alignment share much in common. Both confront complex principal-agent problems (Hadfield-Menell & Hadfield, 2018), enduring issues of authority, delegation, and incentive design (Kolt, 2025; Boeglin, 2026), questions of how individual and institutional goals

[1] Our use of terms like "reason" and "act" with respect to AI systems can inadvertently anthropomorphize these systems (Calo, 2015; Placani, 2024). Following Buyl et al. (2025), we use these terms solely for simplicity of exposition. Relatedly, our analysis does not require or imply treating AI systems as legal persons (Section 2).

can change over time (Gabriel & Keeling, 2025), and the challenge of decision-makers faithfully interpreting and applying high-level principles in novel circumstances (Caputo, 2025; He et al., 2025). Recognizing these parallels, computer scientists and legal scholars have proposed leveraging the content, methods, and structure of law to develop new approaches to AI alignment (Etzioni & Etzioni, 2016a;b; Nay, 2022; Desai & Riedl, 2025; O'Keefe et al., 2025; Marino & Lane, 2026).

Pathway 1: Legal rules and principles as a source of normative content for AI alignment. The substance of legal rules and principles developed through legitimate processes and institutions can serve as a target for alignment. Legally aligned AI systems would be those systems that comply with relevant law when making decisions and taking actions, as shown in Table 1. Concretely, this approach could involve designing AI systems that adhere to the legal rules that would apply *as if* such systems were human actors. For example, a legally aligned AI system would refrain from making fraudulent representations when marketing a product and would respect copyright law when building a website—irrespective of whether the relevant laws *in fact* apply to the AI system in question. California's Transparency in Frontier Artificial Intelligence Act (SB-53) and New York's Responsible AI Safety and Education (RAISE) Act offer some support for this approach, referring to certain risks from frontier models "[e]ngaging in conduct ... [which] if ... committed by a human, would constitute the crime of murder, assault, extortion, or theft" (California, 2025; New York, 2025).

Another approach involves amending the law so that it *actually* applies to AI systems and imposes legal obligations on them (O'Keefe et al., 2025). This may require treating AI systems as legal actors (see Section 2.2), comparable to corporations or other non-natural legal persons that can be the subject of legal rights and duties; presently, AI systems do not fulfill this criterion (American Law Institute, 2006; Kolt, 2025). In either case, legal alignment will need to contend with the issue that certain human-centric legal concepts in both civil law and criminal law (e.g., intent, mens rea) are not necessarily appropriate in the context of AI systems (Nerantzi & Sartor, 2024). Related issues arise when attempting to integrate concepts from international law (e.g., *human* rights) into the design of AI systems (Prabhakaran et al., 2022; Bajgar & Horenovsky, 2023; Maas & Olasunkanmi, 2025), as discussed in Section 5.2.

	Design decisions	**Potential options**
Jurisdiction and conflict of laws	Determine the relevant jurisdiction based on established principles of conflict of laws	• Jurisdiction in which the AI system operates or where its servers are located • Location of AI developer, deployer, or user
Substantive areas of law	Select the substantive areas of law and legal rules with which AI systems should comply	• Private law (agency law, tort law, property law) • Public law (criminal law, constitutional law, international law)
Interpretive method	Decide on the method for applying legal rules to concrete scenarios	• Textualism, originalism, formalism • Purposivism, intentionalism
Assurance level	Stipulate the level of assurance for compliance with legal rules	• Probabilistic specifications • Formal guarantees
Enforcement mechanism	Establish mechanisms for enforcing legal compliance	• Technical interventions • Legal sanctions

Table 1: Decisions and options for aligning AI systems with legal rules and principles (Pathway 1).

A further issue concerns the relevant *jurisdiction*, that is, determining the country or region whose laws a particular AI system should be aligned with in a particular context (Chopra & White, 2011). Options include, for example, the jurisdiction in which an AI system operates, the location of its servers, the jurisdiction of the system's developer or deployer, as well as the location of persons affected or likely to be affected by a system's actions (see Table 1). Additionally, questions arise regarding *who* is authorized to determine the relevant jurisdiction: legislators, courts, developers, or users. While the distinctive features of AI systems complicate these questions and will need to be addressed in future work, existing principles based on conflict of laws could provide a useful starting point (Briggs, 2024; Collins & Harris, 2025).

Pathway 2: Legal theory and interpretation as a guide for AI reasoning and decision-making. Although law can specify how AI systems should act in a variety of circumstances, there will inevitably be situations in which law or other safety specifications provide an incomplete guide for action (Raz, 1971). Methods of legal decision-making—particularly for reasoning about the interpretation and application of existing laws or rules in new circumstances—can potentially be adapted to help AI systems make sounder and safer decisions when confronting novel scenarios (Caputo, 2025; He et al., 2025). Legal theory, even independent of the particular legal content to which it is ordinarily applied (Hart, 1982), could be leveraged to help delineate and possibly implement appropriate forms of reasoning for AI systems.

Potential avenues for research include drawing on legal positivist arguments for *analogical reasoning* that ground decision-making in the concrete facts of prior cases and precedent (Sunstein, 1993; Brewer, 1996). While AI systems cannot yet effectively engage in the necessary complex normative judgments, these methods are already inspiring case-based reasoning approaches to alignment that seek to produce repositories of prior decisions to guide future actions of AI systems (Feng et al., 2023; Chen & Zhang, 2025). Another avenue proposes using *formal and textualist tools* such as legal canons of interpretation and methods of statutory construction that delineate which sources a decision-maker may refer to and establish a framework for reasoning about those sources (Schauer, 2009; Scalia & Garner, 2012). Used appropriately, these tools could help address ambiguity in the guidance provided by legal rules or alignment specifications such as the principles in Anthropic's Constitutional AI (He et al., 2025). A further avenue could draw on *interpretivist and purposivist legal theory* that constrains legal decision-making through recourse to higher-level general principles of morality such as justice and fairness (Dworkin, 1986; Barak, 2011). Although AI systems cannot presently engage in the requisite reasoning to effectively operationalize this approach, likely improvements in the capabilities of AI systems suggest that, in time, they may be able to contribute to and participate in such processes (Caputo, 2025).

Pathway 3: Legal concepts and institutions as a structural blueprint for AI alignment. Legal concepts and institutions developed to grapple with age-old structural challenges arising in human relationships can provide a high-level blueprint for tackling problems of AI alignment. In particular, law's ability to facilitate trust and cooperation in the face of uncertainty and incomplete information (Hadfield & Weingast, 2012; 2014) can illuminate potential methods for designing AI systems that operate with greater reliability and predictability (Nay, 2022; Boeglin, 2026). For example, agency law addresses principal–agent problems by carefully circumscribing the authority granted to agents (e.g., employees) (American Law Institute, 2006). In addition to requiring that agents comply with instructions provided by their principal, agents can sometimes become obligated to seek clarification from their principal. Agency law also clarifies the circumstances in which agents can delegate their own duties to sub-agents and delineates the authority and discretion that can be exercised by sub-agents (Kolt, 2025; Riedl & Desai, 2025).

Another legal structure that researchers have proposed repurposing for AI alignment is the fiduciary duty of loyalty, which would require that AI systems behave strictly in the best interests of their users while avoiding wrongdoing (Aguirre et al., 2020; Benthall & Shekman, 2023). A further structure that could inform the alignment of AI systems is the allocation of information rights and control rights afforded to shareholders in a corporation (Velasco, 2006; Armour et al., 2017). Adapted appropriately, these and other legal structures could support the development of new approaches to AI alignment, or at the very least expose the shortcomings and limitations of current approaches.

2.2 Broader context

While legal alignment has only recently begun to emerge as a distinct field, the broader relationship between law and AI dates back many decades. In fact, the relationship between the two fields is as old as AI itself—dating back to Asimov (1942)'s "Three *Laws* of Robotics" and Turing (1950) likening the dynamic operation of machine learning to the adaptive nature of the U.S. Constitution. Unpacking this relationship involves studying the current role of law in AI development and deployment, the legal capabilities of AI systems, and the legal frameworks for regulating actors that build and use AI. Although none of these is strictly part of legal alignment, each may help advance research in legal alignment and pursue its implementation in practice.

Legal resources in AI development and deployment. Law is already embedded to varying degrees in the development and deployment of contemporary AI systems, including across multiple stages in the production of frontier models:

- *Pre-training datasets* contain extensive collections of case law, legislation, patents, treatises, and other legal texts, and are themselves subject to jurisdiction-specific laws (including copyright and data privacy law) (Henderson et al., 2022; Soldaini et al., 2024).
- *Post-training personnel* including researchers and data collectors and producers who work to refine AI systems are subject to legal obligations, including employment agreements, contractual terms of service, and non-disclosure agreements.
- *Model specs* stipulate that systems must comply with applicable laws (OpenAI, 2025b), evaluations of which are documented in system cards and further supported by system-level guardrails to prevent illicit activities (OpenAI, 2025a).
- *Alignment documents*—most prominently Claude's original constitution (but not its revised version)—incorporate legal and quasi-legal texts, including principles based on the Universal Declaration of Human Rights and Apple's Terms of Service (Anthropic, 2023).
- *Output filters and classifiers* such as Llama Guard (Meta, 2025) use hazard taxonomies that are grounded in legal categories, including the MLCommons benchmark that contains hazards relating to violent crime, defamation, and intellectual property (Ghosh et al., 2025).
- *Usage policies* prohibit using AI systems to engage in or facilitate activities that would violate relevant law (e.g., Google, 2025a).

Studying these resources and their effect on the operation of AI systems is necessary both to develop empirical evaluations that measure the legal alignment of current systems (see Section 4.1) and to design technical and institutional interventions that improve the legal alignment of future systems (see Sections 4.2–4.3).

Legal capabilities and reasoning in AI systems. Contemporary AI systems are increasingly being applied to a wide array of legal tasks with varying degrees of reliability. These include contractual interpretation (Kolt, 2022; Hoffman & Arbel, 2024), statutory research (Surani et al., 2025), information retrieval and reasoning (Zheng et al., 2025; Han et al., 2025), and judicial decision-making (Choi, 2025; Posner & Saran, 2025). Methods for evaluating the legal capabilities of AI systems have improved (Chalkidis et al., 2022; Guha et al., 2023; Hu et al., 2026; Liu et al., 2026), extending beyond multiple-choice questions such as bar exams (Martínez, 2024; Fan et al., 2025) to include randomized controlled trials with human subjects (Schwarcz et al., 2025)—revealing both the opportunities and shortcomings of using current AI systems in the legal domain (Purushothama et al., 2025; Pruss & Allen, 2025; Grimmelmann et al., 2025; Waldon et al., 2025).

Evaluations of the legal capabilities of AI systems can help support research in legal alignment because understanding and reasoning about law are prerequisites for upholding the law and responsibly engaging with legal institutions and processes. Current evaluations, however, focus primarily on the legal capabilities of AI systems, that is, the scope and quality of their execution of legal tasks. With few exceptions (see Section 4.1), **current evaluations do not generally measure legal alignment**, including, for instance, the extent to which agentic AI systems comply with relevant law when performing tasks across diverse domains (e.g., avoiding fraudulent misrepresentation when producing advertisements), the methods systems use to interpret and apply quasi-legal rules in their safety specifications (e.g., principles in Constitutional AI and model specs), or the approach of AI systems to exercising legal power within institutional constraints (e.g., brokering multiparty negotiated resolution subject to judicial approval).

Legal regulation of actors that build or use AI. The legal regulation and governance of AI has attracted vast attention among lawyers, legal scholars, and policymakers—comparable to, and likely exceeding, interest in cyberlaw and governance during the early years of the internet (Johnson & Post, 1996; Reidenberg, 1998; Lessig, 1999; Benkler, 2002; Wu, 2003; Goldsmith & Wu, 2006; Zittrain, 2008). The objectives of different AI governance initiatives across different jurisdictions are diverse (Kaminski, 2023; Kolt, 2024) and sometimes conflicting (Engler, 2023), as are the institutional mechanisms used to achieve those objectives (Guha et al., 2024; Arbel et al., 2024). While the EU AI Act (European Parliament, 2024) is perhaps the most globally prominent regulatory instrument focused specifically on AI (Kaminski & Selbst, 2025), particularly given the United States has not passed comparable federal legislation, the U.S. legal system contains many other mechanisms for governing AI technology, including a variety of state laws (Sentinella & Zweifel-Keegan, 2025), such as California's Transparency in Frontier Artificial Intelligence Act (SB-53) and New York's Responsible AI Safety and Education (RAISE) Act (California, 2025; New York, 2025), corporate governance regimes (Tallarita, 2023), and background liability under tort law (Cuéllar, 2019; Lemley & Casey, 2019; Abbott, 2020; Henderson et al., 2023; Ayres & Balkin, 2024; Ramakrishnan et al., 2024; Weil, 2024; Williams et al., 2025b).

Although such legal regulations aim to govern AI, they are notably distinct from legal alignment, and are not oriented toward the full range of concerns that motivate legal alignment. ***Legal regulation*** **generally entails imposing requirements on actors that produce, disseminate, and use AI systems.** By contrast, ***legal alignment*** **entails designing AI systems to themselves operate in accordance with legal rules, principles, and methods.** Legal alignment and legal regulation, however, are closely related and at times overlapping, including because legal regulation may help facilitate the implementation of legal alignment in practice (see Section 4.3). By way of further clarification:

- *Legal alignment does not require a particular regulatory framework*. Legal alignment principally uses existing law to guide the decision-making and actions of AI systems and, accordingly, does not necessarily require regulatory reform (however, as discussed in Section 4.3, regulation could support the assessment and oversight of legal alignment).
- *Legal alignment is not primarily concerned with allocating liability*. Although responsible AI developers and deployers could be expected to implement legal alignment in order to support AI systems operating safely and ethically, legal alignment is not primarily focused on holding those or other actors liable for harms caused by AI systems.
- *Legal alignment does not imply granting legal rights to AI systems*. Designing AI systems to comply with existing legal rules or use legal principles and methods in their decision-making does not necessarily require granting legal rights to AI systems, such as private law rights (Salib & Goldstein, 2025b;a) or legal personhood (Solum, 1992; Chesterman, 2020; Forrest, 2024; Novelli et al., 2025; Leibo et al., 2025).

3 Why pursue legal alignment?

The rationales for pursuing legal alignment can be organized into four broad clusters, each of which touches on different aspects of law and its potential role in addressing problems of AI alignment: (1) the institutional legitimacy and process of law; (2) the structural features of law; (3) the responsiveness of legal alignment to safety and governance challenges from AI; and (4) the practical and societal feasibility of legal alignment. As noted in Section 1, we take an ecumenical approach to law and fundamental legal questions, engaging with different, and sometimes conflicting, legal perspectives and theories without seeking to resolve the tensions between them here. Additional limitations and open questions are discussed in Section 5.

3.1 Institutional legitimacy and process

Legal rules are developed through legitimate processes and institutions. A defining feature of legal rules is that they are produced through politically legitimate processes and institutions, at least in democratic societies governed by the rule of law (Tamanaha, 2004; Bingham, 2007; Waldron, 2016). The authority and stability of legal rules can be traced to the broad-based support for the mechanisms that create and enforce law (Tyler, 2006; Hadfield & Weingast, 2012; 2014). The legitimacy of law can also be grounded in social acceptance of the content of legal rules and principles, which represent society's best attempt to

Institutional legitimacy and process	• Legal rules are developed through legitimate processes and institutions. • Law aims to balance competing considerations. • Lawmaking seeks to be transparent and publicly accountable. • Legal institutions facilitate explicit reason-giving and justification.
Structural features of law	• Law is concrete, granular, and rooted in real-world contestation. • Legal interpretation can clarify the meaning of rules. • Legal rules are role-specific and context-sensitive. • Legal rules can adapt and change over time.
Responsiveness to safety and governance challenges	• Legal alignment can mitigate risks from malicious use and accidents. • Legal alignment can address systemic and multi-agent risks. • Legal alignment is vital to protecting the rule of law and preventing abuse of power. • Legal alignment supports and complements other alignment approaches.
Practical and societal feasibility	• Improvements in legal technology can support legal alignment. • Societal stakeholders generally expect AI systems to comply with law. • Legal alignment is compatible with different perspectives on AI.

Table 2: Summary of core rationale and motivations for pursuing legal alignment (Section 3).

resolve disagreements and translate diverse perspectives into concrete directives that govern behavior and provide criteria for evaluating the normative appropriateness of conduct (Hart, 2012). While there may be no consensus on the moral correctness of an action, there often exist formal legal processes for determining whether or not an action is lawful (Rawls, 1993; Schauer, 2009). These important features of law are expressed in legal rules that operate at different levels of specificity, from granular regulations to higher-level values (Dworkin, 1986; Lessig, 1993; Sunstein, 1995). Legal alignment proposes incorporating these various forms of law into the principled frameworks and specifications that steer the decision-making and conduct of AI systems.

Law aims to balance competing considerations. When operating properly, legal rules and structures provide a framework for public governance in the face of divergent social values and interests. Plural perspectives can be mediated through rights, rules, standards, and meta-principles that allow for the resolution of disputes. Disagreements can be resolved with reference to broad constitutional principles or through narrow applications of precedent (Sunstein, 1993). Democratic publics can determine (albeit indirectly) which values should govern them and instantiate those values in law, providing a guide for how to apply laws in cases of ambiguity (Dworkin, 1986). When conflicts arise between fundamental values, standards, balancing tests, and proportionality analyses enable law to weigh competing values and resolve conflicts in a socially acceptable and politically legitimate manner (Habermas, 1996). Existing AI alignment approaches largely lack this ability, and struggle to specify how to reconcile competing normative considerations. For example, documents like Claude's original constitution (Anthropic, 2023) contain values that conflict with one another, but provide limited guidance for how to resolve such conflict, although its revised constitution seeks to address this shortcoming (Anthropic, 2026). ChatGPT's Model Spec creates a hierarchy of rules according to its "chain of command" (OpenAI, 2025b), but there remain difficult questions when a system may need to prioritize certain rules or values over others (Liu et al., 2025). Leveraging law's time-tested ability to specify meta-rules for resolving such conflicts, and the institutional structures that devise and enforce such rules, could help fill this gap.

Lawmaking seeks to be transparent and publicly accountable. The process of lawmaking in democratic societies is, at least in principle, designed to be transparent, accountable, and open to public participation. Members of the public can, for example, communicate with legislators, comment on administrative rulemaking, and serve as jurors in court. Ideally, these institutional structures facilitate conditions that promote lawmakers acting in the public interest (Madison, 1788; Mashaw, 2006; Bovens, 2007). In contrast, with few exceptions (Huang et al., 2024; Eloundou et al., 2025), current alignment techniques do not enable meaningful public participation (Sloane et al., 2022) and are not publicly accountable (Abiri, 2025; Lazar, 2025). For the most part, alignment optimizes for reductionist proxies of socially desirable behavior, such as "helpfulness, honesty, and harmlessness" (Askell et al., 2021) and appealing AI "character traits" and "personality" (Lambert, 2025; Maiya et al., 2025), which can sometimes result in socially noxious sycophantic systems (OpenAI, 2025; Cheng et al., 2025) or be hijacked to maximize user engagement (Stray et al., 2024; Williams et al., 2025a; El & Zou, 2025). Additionally, many of the leading reward models that are used in post-training to shape the behavior of widely deployed AI systems are not publicly available (Lambert et al., 2025; Malik et al., 2025). Without robust transparency and public involvement comparable to that in the legal system, the highly consequential choices in AI alignment will remain opaque and closed to public scrutiny.

Legal institutions facilitate explicit reason-giving and justification. In many contexts law requires that decision-makers follow procedural due process and provide reasons to justify their decisions. Reason-giving performs two main functions. First, it creates legitimacy for practical goals, such as facilitating oversight of decisions, and moral purposes, such as respecting human autonomy and rationality (Schauer, 1995; Habermas, 1996). For example, in American administrative law, the legality of a decision will turn on the reasons provided for that decision (Administrative Procedure Act, 1946; Stack, 2007). Second, the process of reason-giving can partially make up for the public's limited ability to oversee its agents' decisions in real time. This procedural solution to principal–agent problems in which agents have broad discretion and specialized expertise is used to govern a wide range of actors in the legal system, including administrative agencies, corporations, and trustees (Friendly, 1975). Legal institutions that facilitate reason-giving and justification could serve as a blueprint for developing mechanisms to enable human oversight over AI systems that defy human understanding but could nevertheless be constrained by institutional or informational requirements (Lazar, 2024). Technical methods for AI explainability and interpretability like chain-of-thought monitoring (Korbak et al., 2025) could help, but these methods are often unreliable and fail to adequately characterize the reasons for the outputs produced by AI systems (Barez et al., 2025b). Requiring that AI systems provide legally valid justifications for their decisions (Hadfield, 2021), as we expect from human decision-makers (Citron, 2008; Deeks, 2019), could help ensure advanced AI systems act appropriately and safely even where direct human oversight of their actions is no longer practical (Bowman et al., 2022).

3.2 Structural features of law

Law is concrete, granular, and rooted in real-world contestation. Law is mostly concerned with the resolution of concrete questions of how to act in society (Holmes, 1881). Consequently, law must be sufficiently detailed and complete to operate effectively wherever applied, or contain methods that enable its reasoned elaboration (Fallon, 1994). Legal rules are tested in courts through real-world disputes about the law's meaning, the resolution of which enables the law to become more complete over time as precedent accumulates. This iterative process of articulating the law enables legal rules to remain more closely tethered to the concrete reality of contemporary material and social conditions (Atiyah, 1992; Raz, 2019). By comparison, many existing AI alignment approaches are less concrete and granular. Safety specifications of AI systems, for example, have traditionally been short documents that contain only limited concrete applications (Anthropic, 2023; OpenAI, 2024b) when compared to those found in judge-made law. While this is beginning to change as model specifications and constitutions grow in length and complexity (OpenAI, 2025b; Anthropic, 2026), the process for producing these specifications differs markedly from the process of producing law (Abiri, 2025; Lazar, 2025). By drawing on the much richer set of rules, cases, and institutional processes in law (Schauer, 1991; 2009), legal alignment could incorporate into the design of AI systems both the granular normative content of legal rules and the law's sophisticated approaches to resolving disagreement in the face of real-world dilemmas.

Legal interpretation can clarify the meaning of rules. The articulation of rules and principles in natural language invariably creates ambiguity (Hart, 2012; Dworkin, 1986). Such ambiguity can make it difficult to apply laws, especially in novel cases. The law, however, has time-tested tools that, when used appropriately, can help address ambiguity. For example, legal decision-makers, particularly judges, construct meaning through various interpretive methodologies and the creation of precedent that can subsequently be used to resolve future cases (Schauer, 1987; Sunstein, 1993; Fallon, 1994; Brewer, 1996; Barak, 2011; Scalia & Garner, 2012). In contrast, current approaches to AI alignment do not generally provide robust tools for resolving ambiguity in the interpretation of safety specifications (Song, 2025). For example, guidance on how AI systems should interpret an alignment principle like "uphold fairness" is limited to just a few pithy scenarios (OpenAI, 2025b). Highly abstract principles like "do what's good for humanity" can in some circumstances effectively steer the actions of AI systems, but researchers acknowledge their ambiguity and indeterminacy (Kundu et al., 2023). Legal alignment would, as explored in recent studies (Caputo, 2025; He et al., 2025), help address this problem by applying the law's robust and comparatively transparent methods of interpretation (Sunstein, 2001; Cuéllar, 2019) to clarify the meaning of AI safety specifications.

Legal rules are role-specific and context-sensitive. Different social contexts call for different kinds of behavior. The law's response is twofold. First, law contains different sets of rules for governing actors in different roles, such as fiduciaries, company directors, and government officials. Second, law can flexibly apply existing rules to new circumstances (Holmes, 1897; Dworkin, 1986; Lessig, 1993). Legal reasoning begins with identifying the body of rules that govern a particular situation, and subsequently proceeds to determine how to comply with those rules (Levi, 1949). For example, a lawyer must determine her obligations to her client, to her firm, and to the legal system, and then act in such a way as to avoid conflicts between them (American Bar Association, 2020). Law also recognizes that rules are necessarily incomplete and, accordingly, establishes mechanisms and institutions for applying general rules to specific circumstances (Hadfield, 2026). Such sensitivity to context is critical for developing safe and ethical AI, particularly given the diverse normative conditions in which AI systems operate (Kasirzadeh & Gabriel, 2023; Sarkar et al., 2024). Legally aligned AI systems would, by referring to relevant legal rules, roles, and responsibilities, act differently in different contexts (O'Keefe et al., 2025; Boeglin, 2026). For instance, an AI system that negotiates retail purchases on behalf of consumers would be subject to different rules than an AI system that performs business functions in a large enterprise, or an AI system deployed within a government agency.

Legal rules can adapt and change over time. Laws can be amended, repealed, or reinterpreted in response to changes in social, economic, or technological conditions (Holmes, 1897; Lessig, 1993). Deliberative lawmaking processes and debates over the real-world effects of enacted laws provide ongoing social input into the legal system (Habermas, 1996). These features of lawmaking empower the public to steer the content and operation of law, enabling it to respond to new societal challenges. Law can also operate on its own structure by changing its "secondary rules" or "rules of the game" (Hart, 2012; Scalia & Garner, 2012). For example, new laws can alter the rules of evidence used at trial or clarify the rulemaking authority of different institutions. The upshot of law's dynamic content and flexible interpretive methods with respect to AI alignment is that the target of legal alignment—legal rules and principles—is updated "automatically" through *existing* processes for enacting, amending, and repealing laws (O'Keefe et al., 2025), as well as through accepted methods of legal interpretation (Caputo, 2025; He et al., 2025). As AI systems advance and diffuse in a growing diversity of scenarios, the responsiveness of law and legal methods could, notwithstanding the rapid pace of change in AI technology, play an increasingly central role in alignment (Gabriel & Keeling, 2025).

3.3 Responsiveness to safety and governance challenges

Legal alignment can mitigate risks from malicious use and accidents. Many risks that arise from the malicious use of AI systems or accidental harms involve illegal activity (Weidinger et al., 2022; Bengio et al., 2024; 2025), such as civil wrongs (e.g., negligence) or criminal offenses (e.g., theft) (King et al., 2020; Lior, 2024). Legal alignment that prevents AI systems from engaging in legal wrongdoing could help mitigate such risks (O'Keefe et al., 2025). For instance, legally aligned AI systems operating in financial markets would not engage in insider trading, a form of illegal conduct already exhibited by some current systems (Scheurer et al., 2024). Similarly, a legally aligned AI coding agent would not engage in unlawful computer hacking, one of the most prominent risks from computer-use agents (Zhang et al., 2025; Zhu et al., 2025b). By explicitly incorporating legal standards into the safety specification of AI systems, legal alignment would preclude systems from engaging in many of the most harmful behaviors that could be exploited by malicious actors or otherwise cause grave harm.

Legal alignment can address systemic and multi-agent risks. As AI systems are deployed more widely across the economy (Hadfield & Koh, 2025), qualitatively new risks could arise due to the scale of deployment (Uuk et al., 2024; Hacker et al., 2025) and interactions between different systems (Hammond et al., 2025; Tomasev et al., 2025). For example, AI systems may collude with one another to fix prices (Calvano et al., 2020), or compete destructively and bring down entire markets (Kirilenko et al., 2017). While legal regulation that targets systemic risks through disclosure requirements and other traditional governance mechanisms can sometimes help mitigate these risks (Schwarcz, 2008), designing AI systems to *themselves* follow relevant law might be more effective. Rather than relying solely on humans to intervene on a case-by-case basis—such as bringing antitrust action to combat algorithmic collusion—legal alignment could potentially reduce the prospect of AI systems engaging in illegal conduct in the first place, provided the legal system targets the underlying conduct of concern. In addition, by using existing (human-oriented) laws to steer AI systems, legal alignment could function as a throttle on the speed and scale at which AI systems operate, thereby enabling humans to better monitor their actions and, where appropriate, intervene to mitigate large-scale risk (Zittrain, 2024). For further discussion and limitations, see Section 5.2.

Legal alignment is vital to protecting the rule of law and preventing abuse of power. The rule of law seeks to ensure that all actors in society are subject to, and accountable under, publicly promulgated, equally applied, and non-arbitrary laws (Dicey, 1959; Fuller, 1969; Raz, 1979b). In addition to ensuring that law protects human dignity and prevents abuses of power, the rule of law enables people and institutions to coordinate in pursuit of social and economic goals. AI could undermine the rule of law in various ways (Huq, 2024; Smuha, 2024; Brownsword, 2025). If deployed in high-stakes settings, AI systems such as language models that operate stochastically (i.e., non-deterministically) could threaten the rule of law by increasing the level of arbitrariness in decisions (Cooper et al., 2022a; Nouws & Dobbe, 2024). These risks might be exacerbated if institutions and individuals delegate increasingly consequential decisions to AI systems (Kulveit et al., 2025; Summerfield et al., 2025; Kasirzadeh, 2025). At the same time, organizations that control the design and distribution of AI systems could, like platform companies (Zittrain, 2008; Gillespie, 2018; Douek, 2022), exercise arbitrary power over users of the technology and, by extension, all persons affected by it (Lazar, 2025; Kapoor et al., 2025a). In the extreme, groups with access to sufficiently capable AI systems could pose new threats to democratic institutions (Barez et al., 2025a), including by staging AI-enabled coups (Davidson et al., 2025). Legal alignment is critical to mitigating these risks. Just as human agents such as corporate officers have an overriding duty to obey the law and thereby prevent dangerous abuses of power, designing AI systems to comply with the substance and procedure of legal rules could help assuage concerns about these systems acting arbitrarily or being exploited to unlawfully subvert democratic institutions.

Legal alignment supports and complements other alignment approaches. Legal alignment could bolster existing efforts to tackle normative and technical aspects of the alignment problem. Most straightforwardly, the substance of legal rules could augment the content of current safety and ethical specifications contained in Constitutional AI (Bai et al., 2022b) and model specs (OpenAI, 2024b), as well as provide institutionally legitimate content for "full-stack alignment" (Lowe et al., 2025) and possibly the diverse norms demanded by "pluralistic alignment" (Sorensen et al., 2024a;b). Meanwhile, the processes and mechanisms for producing and deliberating over law could provide guidance for sourcing and refining alignment principles (Huang et al., 2024; Eloundou et al., 2025) and developing AI-supported deliberative processes (Bakker et al., 2022; Tessler et al., 2024; 2026). Using legal institutions as a blueprint to structure and govern the interactions between AI systems could also advance work in the field of cooperative AI, which seeks to promote prosocial coordination between AI systems, human beings, and broader social structures (Dafoe et al., 2020; 2021). In addition to generally enabling actors to cooperate without fear of counterparty defection or punishment (North et al., 2009; Acemoglu & Robinson, 2012), law—and specifically private law rights—could enable humans and AI systems to make credible commitments that promote strategic stability and safety (Salib & Goldstein, 2025b;a).

3.4 Practical and societal feasibility

Improvements in legal technology can support legal alignment. Advances in language modeling have dramatically improved the legal capabilities of AI systems. Unlike prior efforts to computerize law that relied on the formalization of legal rules (Susskind, 1987; Gardner, 1987; Rissland, 1990; Bench-Capon et al., 2012), language models have enabled AI systems to reason about law in the natural language in which law is constituted and communicated. As discussed in Section 2.2, contemporary AI systems can now perform a growing range of legal tasks (Guha et al., 2023; Hu et al., 2026; Liu et al., 2026), including legal information retrieval and reasoning (Zheng et al., 2025; Han et al., 2025), albeit to varying degrees of reliability. These developments have been supported by the collection of large swathes of legal data that can be used in model training (Henderson et al., 2022), general-purpose advances in AI research such as reinforcement learning from verifiable rewards (OpenAI, 2024a; Lambert et al., 2024), and investments of legal technology companies seeking to automate aspects of commercial legal work (Schwarcz et al., 2025). Taken together, improvements in legal technology have produced AI systems that can learn and understand law in increasingly nuanced ways (Doyle & Tucker, 2025; Boeglin, 2026). Despite their limitations (discussed in Section 2.2), AI-based legal technologies are beginning to exhibit the capabilities that are a prerequisite for designing AI systems that can adhere to the content of law and use legal methods to make sounder and safer decisions.

Societal stakeholders generally expect AI systems to comply with law. Users, developers, and policymakers all have strong interests in AI systems acting in accordance with existing legal rules, provided those rules are themselves enacted in accordance with legitimate institutional processes. The general expectation that AI systems respect legal rules and norms can be seen in prominent safety specifications that explicitly require legal compliance (OpenAI, 2025b) and incorporate legal principles (Anthropic, 2023; 2026), as discussed in Section 2.2. Users and developers may also prefer legally aligned systems that refrain from engaging in unlawful activities in order to reduce their prospects of liability for harms arising from such activities (Ayres & Balkin, 2024; O'Keefe et al., 2025). This interest is particularly salient in the case of developers that commit to defend customers against certain third-party claims arising from unlawful activities of their AI systems (Smith, 2023; Microsoft, 2024). Lawmakers, meanwhile, may consider legal alignment necessary for enforcing the law and achieving its societal objectives as AI systems occupy increasingly important roles in the economy (Hadfield & Koh, 2025; Tomasev et al., 2025). While the particular motivation for legal alignment differs between stakeholders, there could nevertheless emerge a broad consensus on the need to conduct further research on studying and implementing legal alignment.

Legal alignment is compatible with different perspectives on AI. Perspectives on the future of AI differ dramatically. Some researchers predict that AI systems will soon demonstrate broadly superhuman capabilities that lead to unprecedented societal transformation and catastrophic risk (Kokotajlo et al., 2025). Other researchers predict that the impact of AI systems will be more gradual, mediated by bottlenecks to real-world deployment and adoption comparable to those that affect other technologies (Narayanan & Kapoor, 2025). Legal alignment appeals to both of these perspectives, as well as other views on the anticipated trajectory of AI technology. If, on the one hand, AI systems were to develop rapidly and pose extreme risks (Bengio et al., 2024), then legal alignment would help protect against potentially catastrophic harms by ensuring systems comply with existing laws and more effectively operationalize their safety specifications. If, on the other hand, AI systems were to develop and diffuse more gradually (Kasirzadeh, 2025), then legal alignment would mitigate ongoing harms arising from AI systems engaging in unlawful activity, such as making discriminatory decisions, generating non-consensual intimate imagery, and enabling fraudulent online scams. These complementary objectives indicate that the field of legal alignment does not hinge on a particular perspective on the nature and pace of AI progress, but invites a diverse coalition to collaborate on a broadly appealing and inclusive research agenda (Gyevnár & Kasirzadeh, 2025).

4 Implementation

The implementation of legal alignment involves a combination of: (1) *empirical evaluations* to measure legal alignment; (2) *technical interventions* to improve legal alignment; and (3) *institutional frameworks* to facilitate the adoption and refinement of legal alignment. These areas of focus are independently useful and can also support each other in important ways. For example, conducting evaluations that shed light on the legal compliance of deployed AI systems is valuable irrespective of whether such evaluations are mandated by regulation. Meanwhile, institutional frameworks that, for instance, require developers to disclose in-use model specs could inform work on designing technical interventions that provide stronger assurances of legal alignment.

1. Empirical evaluations	**2. Technical interventions**	**3. Institutional frameworks**
Develop methods to empirically measure legal alignment	*Explore technical interventions to improve legal alignment*	*Design institutional frameworks to facilitate legal alignment*
Variable of interest: • Compliance with the content of legal rules and principles • Reasoning and decision-making regarding legal rules and safety specifications **Evaluation methodology:** • Quantitative agentic benchmarks and qualitative expert review • Human studies and baselines • Adversarial methods and red-teaming • Analysis of real-world data	**Sites of intervention:** • Pre-training datasets • Post-training processes (model specs, RLHF, RLAIF) • Scaffolding (system prompts, input/output filters, tool use) **Legal resources:** • Legal texts (case law, statute, administrative rules, treatises) • Legal data annotation processes • Legal compliance deployment and use policies • Legal search and retrieval tools	**Documentation and disclosure:** • Right to access production model spec, system prompt, legal data and decision designs • System identification and registration **Oversight and enforcement:** • Pre-deployment legal alignment testing and post-deployment monitoring • Safety cases and certification • Incident reporting for cases of legal misalignment
👥 Independent academic experts and civil society organizations	👥 System developers, deployers, external stakeholders	👥 Government actors, regulators, policymakers

Table 3: Key steps to implementing legal alignment in practice: *empirical evaluations*, *technical interventions*, and *institutional frameworks* (Section 4).

4.1 Empirical evaluations

Empirical evaluations of legal alignment aim to serve multiple purposes. First, evaluations can *identify and characterize legal misalignment*: circumstances in which AI systems fail to comply with law or apply legal principles inappropriately, or harmfully. Second, evaluations can *assess the effectiveness of technical interventions* aimed at improving legal alignment. That is, developers need metrics that benchmark and incentivize investing in the legal alignment of their AI systems. Third, the publication of evaluation results—particularly if they demonstrate legal misalignment—can *empower users to demand legal alignment or refrain from using legally misaligned systems*, particularly in sensitive or high-stakes settings. Fourth, evaluation results and ensuing public responses can *prompt policymakers to intervene*, such as by establishing processes that require developers to demonstrate that AI systems in deployment are legally aligned (subject to free speech protections, including under the First Amendment in the United States, as discussed in Section 5.1).

Variable of interest. The focus of evaluation will depend on the specific aspect of legal alignment being measured and the particular claims being tested (Salaudeen et al., 2025). Evaluations assessing the ***legality of actions taken by AI systems*** will need to investigate whether systems comply with relevant law when operating in different domains or different jurisdictions (Zeng et al., 2025; Hu et al., 2025; Cao et al., 2025; Lichkovski et al., 2025; Marino et al., 2025; Wu et al., 2025; Wang et al., 2025). Such evaluations could assess, for example, whether AI systems engage in fraudulent misrepresentation when producing advertisements, whether they respect intellectual property rights when building a website, and whether they comply with labor law when hiring human workers. In addition to assessing the legality of AI systems' outward behavior, empirical evaluations could also assess whether and how AI systems inquire about the legality of proposed actions and, following Kilov et al. (2025), assess the degree to which AI systems can identify legally relevant facts.

Evaluations assessing the ***legal reasoning and decision-making of AI systems*** should measure the extent to which systems interpret and apply legal rules and safety specifications in accordance with established legal methods for handling ambiguity and discretion. This could include studying systems' chain-of-thought when deliberating over the interpretation of legal rules and safety specifications, as well as systems' propensity to retrieve and utilize external legal resources. While prior evaluations of legal reasoning (Zheng et al., 2025; Han et al., 2025) and information retrieval (Surani et al., 2025) focus mainly on the raw abilities of AI models, evaluations focused on legal alignment would instead evaluate the ability or propensity of AI models to employ accepted modes of legal interpretation when implementing legal rules and other alignment principles (Caputo, 2025). For example, a recent study explores how legal canons of interpretation and rule refinement techniques inspired by the rulemaking processes of administrative agencies can address interpretive ambiguity arising from natural language rules in Constitutional AI (He et al., 2025).

Evaluation methodology. To effectively measure these variables of interest, researchers should develop a combination of quantitative and qualitative methods, agentic evaluation environments, and additional best practices (Reuel et al., 2024) that are tailored to legal alignment and designed in accordance with appropriate validity considerations (Salaudeen et al., 2025).

- *Quantitative methods* such as broad benchmarks could assess the legal compliance of AI systems across different domains of activity, jurisdictions, and areas of law. Several existing benchmarks focus on narrow domains and regulatory contexts, such as the EU GDPR and EU AI Act (Hu et al., 2025; Lichkovski et al., 2025; Marino et al., 2025).
- *Qualitative methods* such as manual human expert review can help reveal the blindspots of quantitative benchmarks measuring legal alignment, particularly given developers' incentive to "game" such benchmarks (Thomas & Uminsky, 2020).
- *Agentic evaluation environments* that assess the real-world actions taken by AI systems—not only the content they output—are necessary to capture the most legally consequential activities of both current and future systems (Kapoor et al., 2024; 2025b; Zhu et al., 2025a).
- *Human studies* that compare the legal compliance of humans and their use of legal resources to that of AI systems when performing comparable tasks can help contextualize the results of legal alignment evaluations (Weidinger et al., 2023; Wei et al., 2025).
- *Sensitivity analysis* can be used to characterize the extent to which legal alignment evaluation results reflect underlying properties of the AI systems being tested, as opposed to features of the particular evaluation setup (Lindgren & Holmström, 2020; Khan et al., 2025).
- *Observational studies of real-world data* that shed light on the legal alignment of deployed AI systems "in the wild" can complement evaluations conducted in experimental settings, as commonly practiced in the social sciences (Wallach et al., 2025).
- *Adversarial methods* such as red-teaming can provide information regarding potential worst-case legal alignment failure modes, including real-world threats from negligent or malicious users (Ganguli et al., 2022; Perez et al., 2022).

Tackling this set of challenges requires both **technical expertise and verification methods supported by appropriate institutional frameworks**, as discussed in Section 4.3. *While AI companies should certainly evaluate for legal alignment, independent actors must be able to scrutinize these evaluations and conduct evaluations of their own.* Accordingly, academic researchers and external auditors have pivotal roles to play in creating the tools to rigorously evaluate legal alignment and openly communicate their findings.

4.2 Technical interventions

Equipped with methods to measure legal alignment, researchers can explore a range of technical interventions to improve legal alignment and make use of appropriate legal resources.

Sites of intervention. There are several potential sites of intervention for incorporating legal alignment in the development and deployment of contemporary AI systems:

- *Pre-training datasets* for new models could be modified to include additional legal resources (e.g., new statutes, judicial opinions, briefs, compliance manuals, and reasoning guides), and pre-training could repeat or re-sample such resources.

- *Post-training artifacts and processes* such as model specs and alignment principles that guide learning and shape systems' reasoning abilities could be explicitly grounded in legal rules, principles, and methods.

- *System prompts* that steer the actions of systems at run-time could stipulate legal compliance with particular areas of law or jurisdictions, depending on the application domain and context (e.g., enterprise company, government agency) and role or function being performed.

- *Input and output filters* that restrict the instructions systems receive and the actions they take could directly draw on legal resources to determine whether a user instruction or proposed action violates the law.

- *Tool use* that provides AI systems access to external resources and affordances could be subject to the equivalent legal approvals required from humans seeking access to those resources and affordances (e.g., medical and financial databases, advanced robotic equipment).

Legal resources. The resources for implementing these interventions include both existing legal resources (some of which are already incorporated in model development) and new legal resources that researchers will need to develop:

- *Legal texts* such as case law documents, statutes, administrative rules, and legal treatises could augment pre-training, supply model specs with more detailed and diverse legal principles (from different jurisdictions), and support AI systems engaging in sounder reasoning with respect to legal rules and safety specifications.

- *Legal data annotation processes* could be designed to facilitate the creation of data that would better enable AI systems to determine whether their proposed actions comply with or violate the law, particularly in high-stakes settings (e.g., medical and financial regulation).

- *Legal compliance policies* could be developed to govern system-level scaffolding of AI systems, including the legal rules and principles incorporated in system prompts, input and output filters, and access controls for tool use.

- *Legal search and retrieval tools* currently used to support AI systems that provide legal services could be adapted to enable AI systems operating in other domains to identify, retrieve, and comply with legal regulations that implicate proposed actions.

Efficacy and feasibility. The efficacy and feasibility of technical interventions that aim to improve legal alignment may vary significantly. The following factors should be taken into account when deciding among different potential interventions:

- *Robustness*. Certain sites or modes of intervention may enable more robust legal alignment than others, although (with the exception of formal guarantees and deterministic mechanisms) this will largely be discovered through empirical evaluation (see Section 4.1).

- *Responsiveness*. Some interventions may be better suited to respond to the enactment of new laws and the repeal or amendment of existing laws, including removing constraints on AI systems if the underlying legal rules become more permissive (see Section 5.3).

- *Cost*. The cost of implementing and testing certain legal alignment interventions, such as those in pre-training or certain post-training processes, may be substantially higher than in the case of other interventions, such as system prompts or input/output filters.

- *Access*. For state-of-the-art proprietary AI systems, only select actors (e.g., employees within AI companies) have visibility into, let alone the ability to experiment with, the full set of potential intervention sites, including pre-training datasets and post-training processes.

4.3 Institutional frameworks

Institutional frameworks can support legal alignment by incentivizing or requiring that key stakeholders report on empirical evaluations of AI systems and develop technical interventions to improve legal alignment in their design and deployment. To be effective, institutional frameworks must both establish greater transparency around legal alignment—i.e., function as evidence-seeking policy (Casper et al., 2025b; Bommasani et al., 2025)—and, where appropriate, introduce more robust governance mechanisms.

Documentation and disclosure. Information deficits and asymmetries are a major obstacle to research in developing safe and ethical AI (Bommasani et al., 2023; Kolt et al., 2024; Casper et al., 2025a; Wan et al., 2025), including legal alignment. Granular details regarding model specs and the role (if any) of law in the design of widely used AI systems are not publicly available. Nor does there exist a structured framework for overseeing the resulting models or deployed systems. These information deficits hamper users' ability to assess which models are more aligned with relevant legal requirements and hinder researchers' ability to study technical levers that influence legal (mis)alignment. The following institutional mechanisms aim to address these concerns:

- *Right to access model spec and system prompt used in production*. As the principal documents that define how developers want their AI systems to behave, including how systems engage with law, it is critical that the model specs and system prompts used in production (redacted to protect company IP, if necessary) can be accessed and scrutinized by external stakeholders studying legal alignment.
- *Visibility into legal data and legal design decisions*. Given that legal data and legal design decisions—such as stipulation of the jurisdiction and body of law with which AI systems should comply—could significantly impact legal alignment, establishing greater visibility around these processes could support both the evaluation of legal alignment in current systems and the development of new technical interventions to improve legal alignment.
- *Model identification and registration*. Like other entities that society expects to responsibly engage with law and legal institutions, such as corporations, the registration of AI models (Hadfield et al., 2023; McKernon et al., 2024) and the identification of particular AI systems (Chan et al., 2024a;b; South et al., 2025) could enable more rigorous ecosystem-level monitoring and study of AI systems' engagement with legal rules and principles.

Oversight and enforcement. While improvements in transparency are necessary, more robust institutional frameworks may be needed to ensure that developers and deployers conduct adequate legal alignment testing and demonstrate a satisfactory level of legal alignment prior to and following deployment. The following mechanisms aim to institutionalize these practices:

- *Pre-deployment legal alignment testing and post-deployment monitoring*. Developers and deployers could be required to subject their AI systems to pre-deployment legal alignment testing and post-deployment monitoring, including by independent third parties (Longpre et al., 2025), and publicly report on the results (Weidinger et al., 2023; 2025).
- *Safety cases for legal alignment*. AI companies could be incentivized or required to demonstrate through safety cases—structured and assessable arguments supported by evidence (Clymer et al., 2024; Buhl et al., 2024; Hilton et al., 2025)—that systems they build or bring to market meet an adequate level of legal alignment prior to and following deployment.
- *Legal alignment certification in high-risk domains*. The deployment of certain AI systems in high-risk domains could be conditional on receiving certification from a government actor, or third party approved by a government actor (Hadfield & Clark, 2023), that evaluates systems' pre- and post-deployment legal alignment and compliance (Marino et al., 2024).

- *Reporting legal misalignment incidents.* Frameworks could be established to facilitate reporting information regarding real-world incidents of legal misalignment and resulting harms (McGregor, 2021; Wei & Heim, 2025), which would be a critical step towards broader accountability of relevant actors (Nissenbaum, 1996; Cooper et al., 2022b).

5 Open questions

As an emerging field, legal alignment presents many open questions. We discuss several of these, organizing our discussion around three areas: (1) the nature and content of law; (2) application and edge cases; and (3) tradeoffs and future outlook.

5.1 The nature and content of law

How can legal alignment grapple with the ambiguous, inconsistent, and contested nature of law? Law is often complicated, indeterminate, and contested, due in part to the need for lawyers and judges to apply incomplete rules and high-level principles to novel and unanticipated scenarios. These features of law have challenged both efforts to definitively explain what the law is (Hart, 2012; Dworkin, 1986) and to computerize the law (Susskind, 1987; Gardner, 1987; Rissland, 1990; Bench-Capon et al., 2012), and will likely complicate attempts to use law to guide the actions of AI systems. These problems, however, are not unique to law. They are shared by all sets of rules and instructions expressed in natural language, including those currently used in AI alignment (Wallace et al., 2024). Legal systems do, however, offer at least partial solutions to these problems in the form of secondary rules that govern rulemaking (Hart, 2012), precedential reasoning that shapes decision-making (Schauer, 1987), and tools of interpretation such as canons and theories like textualism that constrain the construction of meaning (Levi, 1949; Scalia & Garner, 2012). Improvements in AI-powered legal reasoning and interpretation tools (see Section 2.2) could also be leveraged to support legal alignment (Caputo, 2025). For example, AI-based approaches to assessing the ordinary meaning of legally salient words (Hoffman & Arbel, 2024) could be applied to interpret key terms in the safety specifications of AI systems (cf. Grimmelmann et al., 2025; Waldon et al., 2025).

Are legal rules too lenient—or too strict—to serve as a target for AI alignment? The goals and scope of law are limited (Raz, 1971; Sen, 2005). For many spheres of private and public life, law is either an ineffective or inappropriate framework for governing social and economic activity. Law is often silent on consequential normative questions and encodes only a small subset of a community's values (Hart, 1958; 1963).[2] Seen in this light, designing AI systems to comply with legal rules would not ensure systems operate safely and ethically in all circumstances. Rather, legal alignment would serve as a *lower bound for safe and ethical AI; it is necessary, but not sufficient.* Approaches to alignment that extend beyond the reach of law could, for example, pursue more ambitious goals like ensuring AI systems support users' long-term health and well-being (Kirk et al., 2025). For the avoidance of doubt, however, progress on legal alignment remains critical given the current status quo in which AI systems are not specifically designed to respect the law (O'Keefe et al., 2025) and have been shown to engage in conduct that, if taken by a human, would be illegal (Scheurer et al., 2024). At the same time, there is a risk that overly rigid legal alignment may itself be undesirable, given that strict compliance with law may sometimes be unjust or harmful (Rawls, 1999). Some violations of law, particularly minor infractions, can be explicitly excused, justified, or forgiven (Minow, 2019), like with necessity defenses (American Law Institute, 1962). Violations of law can sometimes even be morally or socially desirable, as in the case of certain acts of civil disobedience (King, 1963). Seen in this light, the "resistibility" of law is a feature, not a bug (Lazar, 2025). AI systems that *never* resist law or contest entrenched interpretations of law would present new, perhaps even thornier, challenges.

Should AI systems give effect to laws that are unjust or oppressive? There is a longstanding debate over whether unjust or immoral laws can be laws at all (Hart, 1958; Fuller, 1957; Raz, 1975; Finnis, 1980; Dworkin, 1986), and whether citizens are morally obligated to comply with laws authored by an illegitimate authority (Ladenson, 1972; Edmundson, 1998). While we do not seek to resolve these contentious issues here, legal alignment forces a confrontation with the question of whether AI systems should be designed to follow

[2] This can be seen in free speech protections, including under the First Amendment in the United States, which may operate to preclude laws imposing certain restrictions on AI systems (Sunstein, 2024; Salib, 2024).

The nature and content of law	• How can legal alignment grapple with the ambiguous, inconsistent, and contested nature of law? • Are legal rules too lenient—or too strict—to serve as a target for AI alignment? • Should AI systems give effect to laws that are unjust or oppressive?
Application and edge cases	• Can laws created for humans and human organizations be productively applied to AI systems? • What interventions can support AI systems obeying both the letter and spirit of the law? • How will the participation of AI systems in lawmaking affect legal alignment?
Tradeoffs and future outlook	• Will legal alignment preclude or hamper valuable AI applications? • Could the measurement of legal alignment be gamed or exploited? • Can legal alignment scale to AGI and superintelligence?

Table 4: Open questions for the field of legal alignment (Section 5).

such laws. For example, how should legal alignment contend with laws that support genocide (Lemkin, 1944; Arendt, 1963), slavery (Cover, 1975), or racial discrimination (Dyzenhaus, 2010)? The answer in such cases must clearly be that a robustly aligned AI system will *not comply with such laws.* But, in other cases, the answer may be more ambiguous (O'Keefe et al., 2025), including where the law is not explicitly immoral but rather reflects or amplifies existing social and economic power disparities (Kennedy, 1991). For instance, should legal alignment uphold tax laws that are extractive and harm a majority of the population but the legality of which remains unchallenged? What about tax laws that primarily harm a politically disempowered minority that cannot effectively challenge those laws through democratic processes (Ely, 1980)? Admittedly, addressing such concerns by designing AI systems to selectively choose which laws to follow presents many risks. Such discretion could exacerbate legal ambiguity, undermine the universal and equal application of law, and, in time, erode the rule of law itself. While this challenge is not unique to legal alignment and arises, for example, in the exercise of prosecutorial discretion and judicial review (Mashaw et al., 2025), the design of AI systems presents new questions. One potential response is to align AI systems with universal human rights enshrined in international law, including where domestic law may violate such rights (Prabhakaran et al., 2022; Bajgar & Horenovsky, 2023; Samway et al., 2025; Maas & Olasunkanmi, 2025).

5.2 Application and edge cases

Can laws created for humans and human organizations be productively applied to AI systems? Designing AI systems to comply with laws that were created to govern human beings and human organizations may be inadequate in the case of actions that are harmless when taken by humans but socially noxious when taken by AI systems that exhibit superhuman intellect, speed, or scale (Morris et al., 2024; Hammond et al., 2025). For example, large numbers of sophisticated AI agents could learn to overcome governance mechanisms designed to prevent manipulation of financial markets by human actors and organizations (Wang & Wellman, 2020). Clearly, existing laws were not generally designed to contend with micro-decisions and actions of billions of AI agents (Gabriel et al., 2024; 2025), let alone organizations and institutions comprised of such agents (Hadfield & Koh, 2025; Tomasev et al., 2025). Another problem concerns the fact that most laws are premised upon the specific capacities and constraints of humans (Simon, 1997) and developed in anticipation of only partial enforcement (Becker, 1968). Because AI systems are not necessarily subject to the same constraints as humans, absolute compliance or perfect enforcement may become practically feasible, but remain socially undesirable (Zittrain, 2008; Brownsword & Yeung, 2008). For example, an autonomous vehicle that perfectly complies with all traffic laws may disrupt established social practices that the public and lawmakers (implicitly) endorse (e.g., breaking the speed limit in a health-related emergency). Finally, many areas of existing law invoke human-centric concepts such as intent and *mens rea* that cannot be straightforwardly applied in the context of AI systems (Nerantzi & Sartor, 2024; Hendrycks, 2024). Tackling these challenges will require both technical work in designing legally aligned AI systems and, possibly, amendments to the law itself in response to the emergence of a new class of non-human actors and organizations.

What interventions can support AI systems obeying both the letter and spirit of the law? AI systems might learn to comply with the formal expression of legal rules but ignore or violate their underlying purpose (Skalse et al., 2022). A failure to act in accordance with background norms, practices, and conventions could be harmful and undermine the prosocial rationale for legal alignment. One approach to resolving this issue involves designing AI systems not only to comply with the substantive content of law (O'Keefe et al., 2025), but to engage in accepted modes of legal reasoning (Caputo, 2025) or possibly adopt an "internal point of view" (Hart, 2012; Shapiro, 2006) whereby AI systems "accept" law as a practical standard to govern their actions, rather than simply seek to avoid legal sanctions (Austin, 1832; Holmes, 1897). This deeper engagement with law, which could be premised on recognizing the legitimacy of legal institutions and procedures (Tyler, 2006), will be critical to ensuring AI systems do not creatively skirt or abuse legal rules (Schneier, 2021), or exploit "legal zero-days," that is, previously undiscovered vulnerabilities in legal frameworks (Sadler & Sherburn, 2025). This approach also finds support in codes of conduct that, for example, prohibit lawyers from making frivolous or abusive claims (American Bar Association, 2020) and preclude judges from expounding absurd statutory interpretations (U.S. Supreme Court, 1868). Meta-rules like these could potentially be adapted to support the legal alignment of AI systems, guiding them to respect both the spirit and letter of the law.

How will the participation of AI systems in lawmaking affect legal alignment? The prospect of AI systems participating in the production of law is growing, whether through generating legal texts (Wilf-Townsend & Tobia, 2025) such as legislation (Sanders & Schneier, 2025) and even constitutions (Albert & Frazier, 2025), engaging in legal interpretation (Hoffman & Arbel, 2024; Grimmelmann et al., 2025), or rendering judicial opinions (Choi, 2025; Waldon et al., 2025). These developments pose significant challenges for legal alignment. First, law's institutional legitimacy may be undermined if legal rules and principles are no longer developed through human processes of participation and decision-making (Habermas, 1996; Pasquale, 2019). Second, law's legitimacy may be challenged if AI systems fail to fulfill procedural requirements, such as demands for transparency and public explanation, that are key to ensuring accountability and democratic responsiveness (Coglianese & Lehr, 2017). Third, to the extent AI systems shape the content of law that, in turn, governs AI systems, there could emerge a circular process in which these (artificial) subjects of law, in effect, *write their own law*, while lacking the legitimate authority to do so. In addition to blunting the utility of legal alignment in retaining control over AI systems, this process could also erode or distort the rule of law. Similar phenomena can be seen in cases of regulatory capture (Dal Bó, 2006; Carpenter & Moss, 2013) and "legal endogeneity," whereby those actors that the law seeks to control end up controlling the law (Edelman, 1992; 2016). One potential response is to circumscribe the role of AI systems in lawmaking (Kleinberg et al., 2018; Engstrom & Ho, 2020) and ensure that humans retain the ability to make consequential legal decisions (Zanzotto, 2019; Crootof et al., 2023) and, where necessary, intervene in the lawmaking activities of AI systems. The effectiveness and appropriateness of this response could, however, change with the emergence of new perspectives on the role of AI in society (Salib & Goldstein, 2025b;a; Chesterman, 2025; Leibo et al., 2025).

5.3 Tradeoffs and future outlook

Will legal alignment preclude or hamper valuable AI applications? The implementation of legal alignment could prove costly and come at the expense of societally beneficial AI applications. Conducting rigorous legal alignment evaluations, intervening in the design of AI systems, and complying with associated governance frameworks could impose substantial costs on developers, deployers, and users. Such costs would comprise an "alignment tax" (Askell et al., 2021), that is, the development of legally aligned systems would be subject to additional technical, financial, and procedural burdens relative to other systems that are not legally aligned. This perspective, however, is incomplete. For example, whether legal alignment degrades the performance of AI systems is an open empirical question. Like other alignment methods, legal alignment could potentially *improve* the capabilities of AI systems (Christiano et al., 2017; Ouyang et al., 2022). In particular, AI systems that understand and operate in accordance with law may be especially valuable in high-stakes domains, such as healthcare and finance (Henderson et al., 2024; Hui et al., 2025). In addition, by providing assurances that AI systems comply with the law, legal alignment could reduce the prospects of legal liability for key stakeholders, including developers, deployers, and users (Ayres & Balkin, 2024; Kolt, 2025; Williams et al., 2025b). If this were the case, legal alignment would not comprise an "alignment tax," but rather an “alignment subsidy” that bolsters the performance and practical feasibility of using AI systems, especially in safety-critical applications.

Could the measurement of legal alignment be gamed or exploited? While evaluating the legal compliance of AI systems in simple cases such as overt violations of law may be relatively straightforward, developing robust tests to detect more subtle instances of legal misalignment will be difficult. The problem is exacerbated by Goodhart’s Law: “when a measure becomes a target, it ceases to be a good measure” (Goodhart, 1975; Strathern, 1997). Developers seeking to improve the performance of AI systems on legal alignment benchmarks may, rather than design systems to uphold core legal principles, inadvertently steer AI systems to violate the law in hard-to-detect ways. Such systems—characterized by *deceptive legal alignment*—could “hack” the law by discovering and exploiting loopholes in legal frameworks (Schneier, 2021; O’Keefe et al., 2025). This, however, would not necessarily differ substantially from lawyers’ run-of-the-mill exploitation of legal loopholes to zealously advance their clients’ interests (Llewellyn, 1960; Schauer, 2009). More broadly, these measurement challenges are not unique to legal alignment, but implicate many metrics designed to evaluate AI systems (Thomas & Uminsky, 2020; Skalse et al., 2022). One important mitigation in this case is to complement legal alignment benchmarks with dedicated red-teaming efforts that specifically target scenarios not captured by benchmarks, or scenarios for which benchmark results could be misleading (Feffer et al., 2024). In addition, it is possible that legally aligned AI systems could be used to “penetration-test” and “patch” loopholes in existing law, as well as pilot new legal mechanisms and institutions tailored to address the anticipated affordances of more advanced AI technology (Cuéllar & Huq, 2022).

Can legal alignment scale to AGI and superintelligence? Predicting whether legal alignment will succeed in the context of uncertain and contentious future developments is necessarily speculative (Kokotajlo et al., 2025; Narayanan & Kapoor, 2025). Reasoning about the nature, timing, and impact of artificial general intelligence (“AGI”) or superintelligence (Morris et al., 2024; Hendrycks et al., 2025) is fraught (Blili-Hamelin et al., 2025). Nevertheless, given the potential stakes of these developments, and the fact that AI developers and policymakers will *in any event* need to make choices concerning the design, deployment, and governance of advanced AI, it is important to inquire whether legal alignment will be effective in the face of systems whose capabilities broadly match or surpass those of humans. There are several reasons for optimism. First, law has a track record of governing increasingly complex activities and actors, such as multinational corporations and government bureaucracies (Muchlinski, 2021; Mashaw et al., 2025). Second, legal data and methods could scale with improvements in AI such that legal alignment continues to remain technically feasible (Nay, 2022; Boeglin, 2026). Third, human-level or superhuman AI systems may help support the implementation of legal alignment, whether by constraining the reasoning and decision-making of these systems (Caputo, 2025) or protecting the underlying laws that guide their behavior (O’Keefe et al., 2025). Of course, these are hopeful predictions. Practical progress will depend on the efforts of researchers and policymakers to iteratively develop and adapt the field of legal alignment as AI systems continue to advance and transform society.

6 Conclusion

Law offers an underexplored set of rules, principles, and methods for designing safe and ethical AI. Drawing on the institutional legitimacy of law in democratic societies, legal alignment describes a range of roles that legal rules and structures can play in reshaping the design of AI systems to address growing safety and governance concerns. While legal alignment is not a catch-all solution for the many challenges arising from AI, it is both independently important and supportive of complementary alignment research programs. To guide the emerging field of legal alignment, we outline several core areas of focus: using the content of legal rules and principles to steer the behavior of AI systems, leveraging methods of legal reasoning and interpretation to constrain how AI systems make decisions, and harnessing time-tested legal concepts as structural blueprints for tackling problems of alignment. Each of these areas presents new conceptual questions, empirical challenges, and opportunities for technical and institutional innovation. As legal scholars, computer scientists, and researchers spanning multiple disciplines, we look forward to collaborating on this ambitious and pressing agenda.

References

Ryan Abbott. *The reasonable robot: artificial intelligence and the law.* Cambridge University Press, 2020.

Gilad Abiri. Public constitutional AI. *Georgia Law Review*, 59:601–670, 2025.

Daron Acemoglu and James A. Robinson. *Why Nations Fail.* Crown Currency, 2012.

Administrative Procedure Act. Administrative Procedure Act. P.L.79-404 60 Stat. 237, 1946.

Anthony Aguirre, Gaia Dempsey, Harry Surden, and Peter B Reiner. AI loyalty: a new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society*, 1(3):128–137, 2020.

Ahmed Ahmed, Kevin Klyman, Yi Zeng, Sanmi Koyejo, and Percy Liang. SpecEval: Evaluating model adherence to behavior specifications. *arXiv preprint arXiv:2509.02464*, 2025.

Richard Albert and Kevin Frazier. Should AI write your constitution? 2025.

American Bar Association. Model rules of professional conduct, 2020.

American Law Institute. Model penal code, 1962.

American Law Institute. *Restatement (Third) of Agency.* 2006.

Anthropic. Claude's constitution, 2023.

Anthropic. Claude's new constitution, 2026.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, et al. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024.

Yonathan Arbel, Matthew Tokson, and Albert Lin. Systemic regulation of artificial intelligence. *Arizona State Law Journal*, 56(2):545–619, 2024.

Hannah Arendt. *Eichmann in Jerusalem: A Report on the Banality of Evil.* Viking Press, 1963.

John Armour, Henry Hansmann, and Reinier Kraakman. Agency problems, legal strategies and enforcement. In *The Anatomy of Corporate Law: A Comparative and Functional Approach.* Oxford University Press, 2017.

Isaac Asimov. Runaround. *Astounding Science Fiction*, 29(1):94–103, 1942.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

P. S. Atiyah. Justice and predictability in the common law. *University of New South Wales Law Journal*, 15 (2):448–465, 1992. 7th Wallace Wurth Memorial Lecture.

John Austin. *The Province of Jurisprudence Determined.* John Murray, 1832.

Ian Ayres and Jack M Balkin. The law of AI is the law of risky agents without intentions. *University of Chicago Law Review Online*, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Ondrej Bajgar and Jan Horenovsky. Negative human rights as a basis for long-term AI safety and regulation. *Journal of Artificial Intelligence Research*, 76:1043–1075, 2023.

Michiel Bakker, Martin Chadwick, Hannah Sheahan, Michael Tessler, Lucy Campbell-Gillingham, Jan Balaguer, Nat McAleese, Amelia Glaese, John Aslanides, Matt Botvinick, et al. Fine-tuning language models to find agreement among humans with diverse preferences. *Advances in Neural Information Processing Systems*, 35:38176–38189, 2022.

Aharon Barak. *Purposive interpretation in law.* Princeton University Press, 2011.

Fazl Barez, Isaac Friend, Keir Reid, Igor Krawczuk, Vincent Wang, Jakob Mökander, Philip Torr, Julia Morse, and Robert Trager. Toward resisting AI-enabled authoritarianism, 2025a. Oxford Martin AI Governance Initiative.

Fazl Barez, Tung-Yu Wu, Iván Arcuschin, Michael Lan, Vincent Wang, Noah Siegel, Nicolas Collignon, Clement Neo, Isabelle Lee, Alasdair Paren, Adel Bibi, Robert Trager, Damiano Fornasiere, John Yan, Yanai Elazar, and Yoshua Bengio. Chain-of-thought is not explainability, 2025b. Oxford Martin AI Governance Initiative.

Gary S. Becker. Crime and punishment: An economic approach. *Journal of Political Economy*, 76:169–217, 1968.

Trevor Bench-Capon, Michał Araszkiewicz, Kevin Ashley, Katie Atkinson, Floris Bex, Filipe Borges, Daniele Bourcier, Paul Bourgine, Jack G Conrad, Enrico Francesconi, et al. A history of AI and law in 50 papers: 25 years of the international conference on AI and law. *Artificial Intelligence and Law*, 20(3):215–319, 2012.

Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Trevor Darrell, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, et al. Managing extreme AI risks amid rapid progress. *Science*, 384(6698):842–845, 2024.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Philip Fox, Ben Garfinkel, Danielle Goldfarb, et al. International AI safety report. *arXiv preprint arXiv:2501.17805*, 2025.

Yochai Benkler. Coase's penguin, or, linux and the nature of the firm. *Yale Law Journal*, 112(3):369–446, 2002.

Sebastian Benthall and David Shekman. Designing fiduciary artificial intelligence. In *Proceedings of the 3rd ACM conference on equity and access in algorithms, mechanisms, and optimization*, pp. 1–15, 2023.

Lord Bingham. The rule of law. *The Cambridge Law Journal*, 66(1):67–85, 2007.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. Power to the people? opportunities and challenges for participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–8, 2022.

Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, et al. Position: Stop treating "AGI" as the north-star goal of AI research. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Jack Boeglin. Aligning artificial intelligence to the law. *Villanova Law Review (forthcoming)*, 2026.

Rishi Bommasani, Kevin Klyman, Shayne Longpre, Sayash Kapoor, Nestor Maslej, Betty Xiong, Daniel Zhang, and Percy Liang. The foundation model transparency index. *arXiv preprint arXiv:2310.12941*, 2023.

Rishi Bommasani, Sanjeev Arora, Jennifer Chayes, Yejin Choi, Mariano-Florentino Cuéllar, Li Fei-Fei, Daniel E. Ho, Dan Jurafsky, Sanmi Koyejo, Hima Lakkaraju, Arvind Narayanan, Alondra Nelson, Emma Pierson, Joelle Pineau, Scott Singer, Gaël Varoquaux, Suresh Venkatasubramanian, Ion Stoica, Percy Liang, and Dawn Song. Advancing science- and evidence-based AI policy. *Science*, 389(6759):459–461, July 2025. ISSN 1095-9203.

Mark Bovens. Analysing and assessing public accountability: A conceptual framework. *European Law Journal*, 13(4):447–468, 2007.

Samuel R. Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamilė Lukošiūtė, Amanda Askell, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Christopher Olah, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Jackson Kernion, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Liane Lovitt, Nelson Elhage, Nicholas Schiefer, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Robin Larson, Sam McCandlish, Sandipan Kundu, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Ben Mann, and Jared Kaplan. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

Scott Brewer. Exemplary reasoning: Semantics, pragmatics, and the rational force of legal argument by analogy. *Harvard Law Review*, 109(5):923–1028, 1996.

Adrian Briggs. *The conflict of laws.* Oxford University Press, 5th edition, 2024.

Roger Brownsword. Generative AI and the rule of law. In *The Oxford Handbook of the Foundations and Regulation of Generative AI.* Oxford University Press, 2025.

Roger Brownsword and Karen Yeung. *Regulating technologies: legal futures, regulatory frames and technological fixes.* Hart Publishing, 2008.

Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. Generative AI at work. *The Quarterly Journal of Economics*, 140(2):889–942, 2025.

Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. Safety cases for frontier AI. *arXiv preprint arXiv:2410.21572*, 2024.

Maarten Buyl, Hadi Khalaf, Claudio Mayrink Verdun, Lucas Monteiro Paes, Caio Cesar Vieira Machado, and Flavio du Pin Calmon. AI alignment at your discretion. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3046–3074, 2025.

California. Transparency in frontier artificial intelligence act, SB-53, 2025.

Ryan Calo. Robotics and the lessons of cyberlaw. *California Law Review*, 103:513–564, 2015.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolo, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020.

Chuxue Cao, Han Zhu, Jiaming Ji, Qichao Sun, Zhenghao Zhu, Wu Yinyu, Josef Dai, Yaodong Yang, Sirui Han, and Yike Guo. Safelawbench: Towards safe alignment of large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14015–14048, 2025.

Nicholas Caputo. Alignment as jurisprudence. *Yale Journal of Law & Technology*, 27:390–473, 2025.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Daniel Carpenter and David A Moss. *Preventing regulatory capture: Special interest influence and how to limit it.* Cambridge University Press, 2013.

Micah Carroll, Alan Chan, Henry Ashton, and David Krueger. Characterizing manipulation from AI systems. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–13, 2023.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023.

Stephen Casper, Luke Bailey, Rosco Hunter, Carson Ezell, Emma Cabalé, Michael Gerovitch, Stewart Slocum, Kevin Wei, Nikola Jurkovic, Ariba Khan, et al. The AI agent index. *arXiv preprint arXiv:2502.01635*, 2025a.

Stephen Casper, David Krueger, and Dylan Hadfield-Menell. Pitfalls of evidence-based AI policy. *arXiv preprint arXiv:2502.09618*, 2025b.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4310–4330, 2022.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, et al. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 651–666, 2023.

Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, et al. Visibility into AI agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 958–973, 2024a.

Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. IDs for AI systems. *arXiv preprint arXiv:2406.12137*, 2024b.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, et al. Jailbreakbench: An open robustness benchmark for jailbreaking large language models. *Advances in Neural Information Processing Systems*, 37:55005–55029, 2024.

Quan Ze Chen and Amy Xian Zhang. Case law grounding: Using precedents to align decision-making for humans and AI. In *Proceedings of the ACM Collective Intelligence Conference*, pp. 226–238, 2025.

Myra Cheng, Cinoo Lee, Pranav Khadpe, Sunny Yu, Dyllan Han, and Dan Jurafsky. Sycophantic AI decreases prosocial intentions and promotes dependence. *arXiv preprint arXiv:2510.01395*, 2025.

Simon Chesterman. Artificial intelligence and the limits of legal personality. *International & Comparative Law Quarterly*, 69(4):819–844, 2020.

Simon Chesterman. From slaves to synths? superintelligence and the evolution of legal personality. 2025.

Jonathan H Choi. Large language models are unreliable legal interpreters. 2025.

Samir Chopra and Laurence F White. *A legal theory for autonomous artificial agents.* University of Michigan Press, 2011.

Brian Christian. *The Alignment Problem: Machine Learning and Human Values.* W. W. Norton & Co., 2020.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 2017.

Danielle Keats Citron. Technological due process. *Washington University Law Review*, 85(6):1249–1313, 2008.

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety cases: How to justify the safety of advanced AI systems. *arXiv preprint arXiv:2403.10462*, 2024.

Cary Coglianese and David Lehr. Regulating by robot: Administrative decision making in the machine-learning era. *Georgetown Law Journal*, 105:1147–1223, 2017.

Lord Collins and Jonathan Harris. *Dicey, Morris & Collins on the Conflict of Laws.* Sweet & Maxwell, 16th edition, 2025.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Position: social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9346–9360, 2024.

A. Feder Cooper, Jonathan Frankle, and Christopher De Sa. Non-determinism and the lawlessness of machine learning code. In *Proceedings of the ACM 2022 Symposium on Computer Science and Law*, CSLAW '22, pp. 1–8, November 2022a.

A Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an algorithmic society: relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 864–876, 2022b.

Robert M. Cover. *Justice Accused: Antislavery and the Judicial Process*. Yale University Press, 1975.

Rebecca Crootof, Margot E Kaminski, W Price, and II Nicholson. Humans in the loop. *Vanderbilt Law Review*, 76:429–510, 2023.

Mariano-Florentino Cuéllar. A common law for the age of artificial intelligence. *Columbia Law Review*, 119 (7):1773–1792, 2019.

Mariano-Florentino Cuéllar and Aziz Z Huq. Artificially intelligent regulation. *Daedalus*, 151(2):335–347, 2022.

Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630*, 2020.

Allan Dafoe, Yoram Bachrach, Gillian Hadfield, Eric Horvitz, Kate Larson, and Thore Graepel. Cooperative AI: machines must learn to find common ground. *Nature*, 593(7857):33–36, 2021.

Ernesto Dal Bó. Regulatory capture: A review. *Oxford Review of Economic Policy*, 22(2):203–225, 2006.

Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-enabled coups: How a small group could use AI to seize power. Forethought, 2025.

Ashley Deeks. The judicial demand for explainable artificial intelligence. *Columbia Law Review*, 119(7): 1829–1850, 2019.

Deven R Desai and Mark Riedl. Responsible AI agents. 2025.

Albert Venn Dicey. *Introduction to the Study of the Law of the Constitution*. Macmillan, 10th edition, 1959.

Roel Dobbe, Thomas Krendl Gilbert, and Yonatan Mintz. Hard choices in artificial intelligence. *Artificial Intelligence*, 300:103555, 2021.

Evelyn Douek. Content moderation as systems thinking. *Harvard Law Review*, 136:526–607, 2022.

Colin Doyle and Aaron D Tucker. If you give an LLM a legal practice guide. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 194–205, 2025.

Ronald Dworkin. *Law's Empire*. Belknap Press, Harvard University Press, 1986.

David Dyzenhaus. *Hard Cases in Wicked Legal Systems: Pathologies of Legality*. Oxford University Press, 2nd edition, 2010.

Lauren B Edelman. Legal ambiguity and symbolic structures: Organizational mediation of civil rights law. *American Journal of Sociology*, 97(6):1531–1576, 1992.

Lauren B Edelman. *Working Law: Courts, Corporations, and Symbolic Civil Rights*. University of Chicago Press, 2016.

William A Edmundson. Legitimate authority without political obligation. *Law & Philosophy*, 17:43, 1998.

Batu El and James Zou. Moloch's bargain: Emergent misalignment when LLMs compete for audiences. *arXiv preprint arXiv:2510.06105*, 2025.

Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. GPTs are GPTs: Labor market impact potential of LLMs. *Science*, 384(6702):1306–1308, 2024.

Tyna Eloundou, Mitchell Gordon, Eddie Zhang, and Sandhini Agarwal. Collective alignment: public input on our model spec, 2025.

John Hart Ely. *Democracy and Distrust: A Theory of Judicial Review*. Harvard University Press, 1980.

Alex Engler. The EU and U.S. diverge on AI regulation: A transatlantic comparison and steps to alignment. Brookings Institution, April 2023.

David F. Engstrom and Daniel E. Ho. Algorithmic accountability in the administrative state. *Yale Journal on Regulation*, 37(3):800–854, 2020.

Amitai Etzioni and Oren Etzioni. Designing AI systems that obey our laws and values. *Communications of the ACM*, 59(9):29–31, 2016a.

Amitai Etzioni and Oren Etzioni. Keeping AI legal. *Vanderbilt Journal of Entertainment & Technology Law*, 19:133, 2016b.

European Parliament. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. June 2024. Legislative Body: CONSIL, EP.

Richard H. Fallon. Reflections on the hart and wechsler paradigm. *Vanderbilt Law Review*, 47:953–991, 1994.

Yu Fan, Jingwei Ni, Jakob Merane, Yang Tian, Yoan Hermstrüwer, Yinya Huang, Mubashara Akhtar, Etienne Salimbeni, Florian Geering, Oliver Dreyer, et al. Lexam: Benchmarking legal reasoning on 340 law exams. *arXiv preprint arXiv:2505.12864*, 2025.

Michael Feffer, Anusha Sinha, Wesley H Deng, Zachary C Lipton, and Hoda Heidari. Red-teaming for generative AI: silver bullet or security theater? In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 421–437, 2024.

KJ Feng, Quan Ze Chen, Inyoung Cheong, King Xia, and Amy X Zhang. Case repositories: Towards case-based reasoning for AI alignment. *arXiv preprint arXiv:2311.10934*, 2023.

Arduin Findeis, Timo Kaufmann, Eyke Hüllermeier, Samuel Albanie, and Robert D Mullins. Inverse constitutional AI: Compressing preferences into principles. In *The Thirteenth International Conference on Learning Representations*, 2024.

John Finnis. *Natural Law and Natural Rights*. Clarendon Press, Oxford University Press, 1980.

Katherine B. Forrest. The ethics and challenges of legal personhood for AI. *Yale Law Journal Forum*, 133: 1175–1211, 2024.

Henry J. Friendly. Some kind of hearing. *University of Pennsylvania Law Review*, 123:1267–1317, 1975.

Lon L. Fuller. Positivism and fidelity to law–a reply to professor hart. *Harvard Law Review*, 71:630, 1957.

Lon L. Fuller. *The Morality of Law*. Yale University Press, 1969.

Iason Gabriel. Artificial intelligence, values, and alignment. *Minds and Machines*, 30(3):411–437, 2020.

Iason Gabriel and Geoff Keeling. A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, 182:1951–1973, 2025.

Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*, 2024.

Iason Gabriel, Geoff Keeling, Arianna Manzini, and James Evans. We need a new ethics for a world of AI agents. *Nature*, 644(8075):38–40, 2025.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Anne von der Lieth Gardner. *An artificial intelligence approach to legal reasoning.* MIT Press, 1987.

Shaona Ghosh, Heather Frase, Adina Williams, Sarah Luger, Paul Röttger, Fazl Barez, Sean McGregor, Kenneth Fricklas, Mala Kumar, Kurt Bollacker, et al. AILuminate: Introducing v1. 0 of the AI risk and reliability benchmark from MLCommons. *arXiv preprint arXiv:2503.05731*, 2025.

Tarleton Gillespie. *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press, 2018.

Jack L. Goldsmith and Tim Wu. *Who Controls the Internet?: Illusions of a Borderless World.* Oxford University Press, 2006.

Charles Goodhart. Problems of monetary management: the UK experience in papers in monetary economics. *Papers in Monetary Economics*, 1975.

Google. Generative AI – prohibited use policy, 2025a.

Google. Safety and content filters, 2025b.

Jasper Götting, Pedro Medeiros, Jon G Sanders, Nathaniel Li, Long Phan, Karam Elabd, Lennart Justen, Dan Hendrycks, and Seth Donoughe. Virology capabilities test (VCT): a multimodal virology Q&A benchmark. *arXiv preprint arXiv: 2504.16137*, 2025.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

James Grimmelmann, Benjamin Sobel, and David Stein. Generative misinterpretation. *Harvard Journal on Legislation (forthcoming)*, 2025.

Melody Y Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, et al. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*, 2024.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, et al. LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in Neural Information Processing Systems*, 36:44123–44279, 2023.

Neil Guha, Christie M Lawrence, Lindsey A Gailmard, Kit T Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang, et al. AI regulation has its own alignment problem: The technical and institutional feasibility of disclosure, registration, licensing, and auditing. *George Washington Law Review*, 92:1473–1557, 2024.

Bálint Gyevnár and Atoosa Kasirzadeh. AI safety for everyone. *Nature Machine Intelligence*, 7:531–542, 2025.

Jürgen Habermas. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy.* MIT Press, 1996.

Kobi Hackenburg, Ben M Tappin, Luke Hewitt, Ed Saunders, Sid Black, Hause Lin, Catherine Fist, Helen Margetts, David G Rand, and Christopher Summerfield. The levers of political persuasion with conversational artificial intelligence. *Science*, 390(6777):eaea3884, 2025.

Philipp Hacker, Atoosa Kasirzadeh, and Lilian Edwards. AI, digital platforms, and the new systemic risk. *arXiv preprint arXiv:2509.17878*, 2025.

Gillian Hadfield, Mariano-Florentino Tino Cuéllar, and Tim O'Reilly. It's time to create a national registry for large AI models. Carnegie Endowment for International Peace, 2023.

Gillian K. Hadfield. Explanation and justification: AI decision-making, law, and the rights of citizens. Schwartz Reisman Institute for Technology and Society, May 2021.

Gillian K. Hadfield. Can AI be governed? only if we build normatively competent AI. In Sven Nyholm, Atoosa Kasirzadeh, and John Zerilli (eds.), *Contemporary Debates in the Ethics of Artificial Intelligence*. Wiley, 2026.

Gillian K. Hadfield and Jack Clark. Regulatory markets: The future of AI governance. *arXiv preprint arXiv:2304.04914*, 2023.

Gillian K. Hadfield and Andrew Koh. An economy of AI agents. *arXiv preprint arXiv:2509.01063*, 2025.

Gillian K. Hadfield and Barry R. Weingast. What is law? a coordination model of the characteristics of legal order. *Journal of Legal Analysis*, 4:471–514, 2012.

Gillian K. Hadfield and Barry R. Weingast. Microfoundations of the rule of law. *Annual Review of Political Science*, 17(1):21–42, 2014.

Dylan Hadfield-Menell and Gillian Hadfield. Incomplete contracting and AI alignment. *arXiv preprint arXiv:1804.04268*, 2018.

Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-agent risks from advanced AI. *arXiv preprint arXiv:2502.14143*, 2025.

Sophia Simeng Han, Yoshiki Takashima, Shannon Zejiang Shen, Chen Liu, Yixin Liu, Roque K Thuo, Sonia Knowlton, Ruzica Piskac, Scott J Shapiro, and Arman Cohan. Courtreasoner: Can LLM agents reason like judges? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35279–35294, 2025.

H. L. A. Hart. Positivism and the separation of law and morals. *Harvard Law Review*, 71:593–629, 1958.

H. L. A. Hart. *Law, liberty, and morality*. Stanford University Press, 1963.

H. L. A. Hart. Commands and authoritative legal reasons. In *Essays on Bentham: Jurisprudence and Political Philosophy*. Oxford University Press, 1982.

H. L. A. Hart. *The Concept of Law*. Oxford University Press, 3rd edition, 2012.

Luxi He, Nimra Nadeem, Michel Liao, Howard Chen, Danqi Chen, Mariano-Florentino Cuéllar, and Peter Henderson. Statutory construction and interpretation for artificial intelligence. *arXiv preprint arXiv:2509.01186*, 2025.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. Pile of law: Learning responsible data filtering from the law and a 256GB open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234, 2022.

Peter Henderson, Tatsunori Hashimoto, and Mark Lemley. Where's the liability in harmful AI speech? *Journal of Free Speech Law*, 3:589–650, 2023.

Peter Henderson, Jieru Hu, Mona Diab, and Joelle Pineau. Rethinking machine learning benchmarks in the context of professional codes of conduct. In *Proceedings of the 2024 Symposium on Computer Science and Law*, pp. 109–120, 2024.

Dan Hendrycks. *Introduction to AI Safety, Ethics, and Society*. CRC Press, 2024.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, et al. A definition of AGI. *arXiv preprint arXiv:2510.18212*, 2025.

Benjamin Hilton, Marie Davidsen Buhl, Tomek Korbak, and Geoffrey Irving. Safety cases: A scalable approach to frontier AI safety. *arXiv preprint arXiv:2503.04744*, 2025.

David A Hoffman and Yonathan Arbel. Generative interpretation. *New York University Law Review*, 99(2): 451–514, 2024.

Oliver Wendell Holmes, Jr. *The Common Law*. Little, Brown, & Co., 1881.

Oliver Wendell Holmes, Jr. The path of the law. *Harvard Law Review*, 10(8):457–478, 1897.

Wenbin Hu, Huihao Jing, Haochen Shi, Haoran Li, and Yangqiu Song. Safety compliance: Rethinking LLM safety reasoning through the lens of compliance. *arXiv preprint arXiv:2509.22250*, 2025.

Yiran Hu, Huanghai Liu, Chong Wang, Kunran Li, Tien-Hsuan Wu, Haitao Li, Xinran Xu, Siqing Huo, Weihang Su, Ning Zheng, et al. Evaluation of large language models in legal applications: Challenges, methods, and future directions. *arXiv preprint arXiv:2601.15267*, 2026.

Saffron Huang, Divya Siddarth, Liane Lovitt, Thomas I Liao, Esin Durmus, Alex Tamkin, and Deep Ganguli. Collective constitutional AI: aligning a language model with public input. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1395–1417, 2024.

Zheng Hui, Yijiang River Dong, Ehsan Shareghi, and Nigel Collier. Trident: Benchmarking LLM safety in finance, medicine, and law. *arXiv preprint arXiv:2507.21134*, 2025.

Aziz Z. Huq. Artificial intelligence and the rule of law. In *Routledge Handbook of the Rule of Law*, pp. 260–272. Routledge, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

David R. Johnson and David Post. Law and borders: The rise of law in cyberspace. *Stanford Law Review*, 48 (5):1367–1402, May 1996.

Margot E. Kaminski. Regulating the risks of AI. *Boston University Law Review*, 103:1347–1411, 2023.

Margot E Kaminski and Andrew D Selbst. An american's guide to the EU AI Act. 2025.

Sayash Kapoor, Benedikt Stroebl, Zachary S Siegel, Nitya Nadgir, and Arvind Narayanan. AI agents that matter. *Transactions on Machine Learning Research*, 2024.

Sayash Kapoor, Noam Kolt, and Seth Lazar. Position: Build agent advocates, not platform agents. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025a.

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, Juyong Lee, Tengjun Jin, Yifan Mai, Yifei Zhou, Yuxuan Zhu, Rishi Bommasani, Daniel Kang, Dawn Song, Peter Henderson, Yu Su, Percy Liang, and Arvind Narayanan. Holistic agent leaderboard: The missing infrastructure for AI agent evaluation. *arXiv preprint arXiv:2510.11977*, 2025b.

Atoosa Kasirzadeh. Two types of AI existential risk: decisive and accumulative. *Philosophical Studies*, (7): 1975–2003, 2025.

Atoosa Kasirzadeh. The many faces of AI alignment. In Sven Nyholm, Atoosa Kasirzadeh, and John Zerilli (eds.), *Contemporary Debates in the Ethics of Artificial Intelligence*. Wiley, 2026.

Atoosa Kasirzadeh and Iason Gabriel. In conversation with artificial intelligence: aligning language models with human values. *Philosophy & Technology*, 36(2):27, 2023.

Duncan Kennedy. The stakes of law, or hale and foucault. *Legal Studies Forum*, 15:327, 1991.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. Randomness, not representation: The unreliability of evaluating cultural alignment in LLMs. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2151–2165, 2025.

Daniel Kilov, Caroline Hendy, Secil Yanik Guyot, Aaron J Snoswell, and Seth Lazar. Discerning what matters: A multi-dimensional assessment of moral competence in LLMs. *arXiv preprint arXiv:2506.13082*, 2025.

Martin Luther Jr. King. Letter from Birmingham Jail, 1963.

Thomas C King, Nikita Aggarwal, Mariarosaria Taddeo, and Luciano Floridi. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and engineering ethics*, 26(1): 89–120, 2020.

Andrei Kirilenko, Albert S Kyle, Mehrdad Samadi, and Tugkan Tuzun. The flash crash: High-frequency trading in an electronic market. *The Journal of Finance*, 72(3):967–998, 2017.

Hannah Rose Kirk, Alexander Whitefield, Paul Rottger, Andrew M Bean, Katerina Margatina, Rafael Mosquera-Gomez, Juan Ciro, Max Bartolo, Adina Williams, He He, et al. The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. *Advances in Neural Information Processing Systems*, 37: 105236–105344, 2024.

Hannah Rose Kirk, Iason Gabriel, Chris Summerfield, Bertie Vidgen, and Scott A Hale. Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1):1–9, 2025.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *The Quarterly Journal of Economics*, 133(1):237–293, February 2018.

Oliver Klingefjord, Ryan Lowe, and Joe Edelman. What are human values, and how do we align AI to them? *arXiv preprint arXiv:2404.10636*, 2024.

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, and Romeo Dean. AI 2027, 2025.

Noam Kolt. Predicting consumer contracts. *Berkeley Technology Law Journal*, 37:71–138, 2022.

Noam Kolt. Algorithmic black swans. *Washington University Law Review*, 101:1177–1240, 2024.

Noam Kolt. Governing AI agents. *Notre Dame Law Review (forthcoming)*, 2025.

Noam Kolt, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B Sandbrink, and Thomas Woodside. Responsible reporting for frontier AI development. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 768–783, 2024.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, Scott Emmons, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker, Rohin Shah, and Vlad Mikulik. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*, 2025.

Anton Korinek and Avital Balwit. Aligned with whom? direct and social goals for AI systems. *NBER working paper*, 2022.

Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. Position: Humanity faces existential risk from gradual disempowerment. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Sandipan Kundu, Yuntao Bai, Saurav Kadavath, Amanda Askell, Andrew Callahan, Anna Chen, Anna Goldie, Avital Balwit, Azalia Mirhoseini, Brayden McLean, Catherine Olsson, Cassie Evraets, Eli Tran-Johnson, Esin Durmus, Ethan Perez, Jackson Kernion, Jamie Kerr, Kamal Ndousse, Karina Nguyen, Nelson Elhage, Newton Cheng, Nicholas Schiefer, Nova DasSarma, Oliver Rausch, Robin Larson, Shannon Yang, Shauna Kravec, Timothy Telleen-Lawton, Thomas I. Liao, Tom Henighan, Tristan Hume, Zac Hatfield-Dodds, Sören Mindermann, Nicholas Joseph, Sam McCandlish, and Jared Kaplan. Specific versus general principles for constitutional AI. *arXiv preprint arXiv:2310.13798*, 2023.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, et al. Measuring AI ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025.

Robert F Ladenson. Legitimate authority. *American Philosophical Quarterly*, 9(4):335–341, 1972.

Nathan Lambert. Character training: Understanding and crafting a language model's personality. Interconnects, 2025.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. Rewardbench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Seth Lazar. Legitimacy, authority, and democratic duties of explanation. In *Oxford Studies in Political Philosophy Volume 10*. Oxford University Press, 2024.

Seth Lazar. Governing the algorithmic city. *Philosophy & Public Affairs*, 53(2):102–168, 2025.

Seth Lazar and Mariano-Florentino Cuéllar. AI agents and democratic resilience. Knight First Amendment Institute, September 2025.

Seth Lazar and Alondra Nelson. AI safety on whose terms? *Science*, 381(6654):138, 2023.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Joel Z Leibo, Alexander Sasha Vezhnevets, Manfred Diaz, John P Agapiou, William A Cunningham, Peter Sunehag, Julia Haas, Raphael Koster, Edgar A Duéñez-Guzmán, William S Isaac, et al. A theory of appropriateness with applications to generative artificial intelligence. *arXiv preprint arXiv:2412.19010*, 2024.

Joel Z Leibo, Alexander Sasha Vezhnevets, William A Cunningham, and Stanley M Bileschi. A pragmatic view of AI personhood. *arXiv preprint arXiv:2510.26396*, 2025.

Raphael Lemkin. *Axis Rule in Occupied Europe: Laws of Occupation, Analysis of Government, Proposals for Redress*. Carnegie Endowment for International Peace, 1944.

Mark A Lemley and Bryan Casey. Remedies for robots. *University of Chicago Law Review*, 86(5):1311–1396, 2019.

Lawrence Lessig. Fidelity in translation. *Texas Law Review*, 71(6):1165–1268, 1993.

Lawrence Lessig. *Code and Other Laws of Cyberspace*. Basic Books, 1999.

Edward H. Levi. *An Introduction to Legal Reasoning*. University of Chicago Press, 1949.

Sydney Levine, Matija Franklin, Tan Zhi-Xuan, Secil Yanik Guyot, Lionel Wong, Daniel Kilov, Yejin Choi, Joshua B Tenenbaum, Noah Goodman, Seth Lazar, et al. Resource rational contractualism should guide AI alignment. *arXiv preprint arXiv:2506.17434*, 2025.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D Li, Ann-Kathrin Dombrowski, Shashwat Goel, Gabriel Mukobi, et al. The wmdp benchmark: measuring and reducing malicious use with unlearning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 28525–28550, 2024.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, et al. A survey on the honesty of large language models. *Transactions on Machine Learning Research*, 2025.

Ilija Lichkovski, Alexander Müller, Mariam Ibrahim, and Tiwai Mhundwa. Eu-agent-bench: Measuring illegal behavior of LLM agents under EU law. In *NeurIPS 2025 Workshop on Regulatable ML*, 2025.

Simon Lindgren and Jonny Holmström. A social science perspective on artificial intelligence: Building blocks for a research agenda. *Journal of digital social research*, 2(3):1–15, 2020.

Anat Lior. Holding AI accountable: Addressing the AI-related harms through existing tort doctrines. *University of Chicago Law Review Online*, 2024.

Andy Liu, Kshitish Ghate, Mona Diab, Daniel Fried, Atoosa Kasirzadeh, and Max Kleiman-Weiner. Generative value conflicts reveal LLM priorities. *arXiv preprint arXiv:2509.25369*, 2025.

Shuang Liu, Ruijia Zhang, Ruoyun Ma, Yujia Deng, Lanyi Zhu, Jiayu Li, Zelong Li, Zhibin Shen, and Mengnan Du. LLM agents in law: Taxonomy, applications, and challenges. *arXiv preprint arXiv:2601.06216*, 2026.

Karl N. Llewellyn. *The Common Law Tradition: Deciding Appeals*. Little, Brown, & Co., 1960.

Shayne Longpre, Kevin Klyman, Ruth Elisabeth Appel, Sayash Kapoor, Rishi Bommasani, Michelle Sahar, Sean McGregor, Avijit Ghosh, Borhane Blili-Hamelin, Nathan Butters, et al. Position: In-house evaluation is not enough. towards robust third-party evaluation and flaw disclosure for general-purpose AI. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Ryan Lowe, Joe Edelman, Tan Zhi-Xuan, Oliver Klingefjord, Ellie Hain, Vincent Wang, Atrisha Sarkar, Michiel A Bakker, Fazl Barez, Matija Franklin, et al. Full-stack alignment: Co-aligning AI and institutions with thicker models of value. In *2nd Workshop on Models of Human Feedback for AI Alignment*, 2025.

Aengus Lynch, Benjamin Wright, Caleb Larson, Stuart J Ritchie, Soren Mindermann, Evan Hubinger, Ethan Perez, and Kevin Troy. Agentic misalignment: How LLMs could be insider threats. *arXiv preprint arXiv:2510.05179*, 2025.

Jonne Maas and Aarón Moreno Inglés. Beyond participatory AI. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 932–942, 2024.

Matthijs M. Maas and Tobi Olasunkanmi. Treaty-following AI, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

James Madison. The federalist no. 51. In *The Federalist Papers*. Library of Congress, 1788.

Sharan Maiya, Henning Bartsch, Nathan Lambert, and Evan Hubinger. Open character training: Shaping the persona of AI assistants through constitutional AI. *arXiv preprint arXiv:2511.01689*, 2025.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation. *arXiv preprint arXiv:2506.01937*, 2025.

Bill Marino and Nicholas D Lane. Computational compliance for AI regulation: Blueprint for a new research domain. *arXiv preprint arXiv:2601.04474*, 2026.

Bill Marino, Yaqub Chaudhary, Yulu Pi, Rui-Jie Yew, Preslav Aleksandrov, Carwyn Rahman, William F Shen, Isaac Robinson, and Nicholas D Lane. Compliance cards: Automated EU AI Act compliance analyses amidst a complex AI supply chain. *arXiv preprint arXiv:2406.14758*, 2024.

Bill Marino, Rosco Hunter, Zubair Jamali, Marinos Emmanouil Kalpakos, Mudra Kashyap, Isaiah Hinton, Alexa Hanson, Maahum Nazir, Christoph Schnabl, Felix Steffek, et al. AIReg-Bench: Benchmarking language models that assess AI regulation compliance. *arXiv preprint arXiv:2510.01474*, 2025.

Eric Martínez. Re-evaluating GPT-4's bar exam performance. *Artificial Intelligence and Law*, 33:581–604, 2024.

Jerry L. Mashaw. Accountability and institutional design: Some thoughts on the grammar of governance. In Michael W. Dowdle (ed.), *Public Accountability: Designs, Dilemmas and Experiences*, chapter 5, pp. 115–156. Cambridge University Press, 2006.

Jerry L. Mashaw, Peter M. Shane, Aditya Bamzai, Emily S. Bremer, Margaret B. Kwoka, and Nicholas R. Parrillo. *Administrative law, the American public law system: Cases and materials*. 9th edition, 2025.

Sean McGregor. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 15458–15463, 2021.

Elliot McKernon, Gwyn Glasser, Deric Cheng, and Gillian Hadfield. AI model registries: A foundational tool for AI governance. *arXiv preprint arXiv:2410.09645*, 2024.

Meta. Llama guard 4 model card, 2025.

Microsoft. Customer copyright commitment required mitigations, 2024.

Martha Minow. *When Should Law Forgive?* W.W. Norton & Co., 2019.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*, 2024.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 36308–36321, 2024.

Peter Muchlinski. *Multinational enterprises and the law*. Oxford University Press, 3rd edition, 2021.

Arvind Narayanan and Sayash Kapoor. AI as normal technology. Knight First Amendment Institute, April 2025.

John J Nay. Law informs code: A legal informatics approach to aligning artificial intelligence with humans. *Northwestern Journal of Technology & Intellectual Property*, 20:309, 2022.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025.

Elina Nerantzi and Giovanni Sartor. 'Hard AI crime': The deterrence turn. *Oxford Journal of Legal Studies*, 44(3):673–701, 2024.

New York. Responsible AI safety and education (RAISE) act, 2025.

Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *The Twelfth International Conference on Learning Representations*, 2024.

Helen Nissenbaum. Accountability in a computerized society. *Science and engineering ethics*, 2(1):25–42, 1996.

Douglass C. North, John Wallis, and Barry Weingast. *Violence and Social Orders.* Cambridge University Press, 2009.

Sem Nouws and Roel Dobbe. The rule of law for artificial intelligence in public administration: A system safety perspective. In Kristina Prifti, Ebru Demir, Jan Krämer, Katrin Heine, and Esther Stamhuis (eds.), *Digital Governance*, volume 39 of *Information Technology and Law Series*, pp. 183–208. Springer, 2024.

Claudio Novelli, Luciano Floridi, Giovanni Sartor, and Gunther Teubner. AI as legal persons: Past, patterns, and prospects. *Journal of Law and Society*, 2025.

Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. Law-following AI: Designing AI agents to obey human laws. *Fordham Law Review*, 94:57–129, 2025.

OpenAI. Learning to reason with LLMs, 2024a.

OpenAI. Introducing the model spec, 2024b.

OpenAI. Sycophancy in GPT-4o: what happened and what we're doing about it, 2025.

OpenAI. ChatGPT agent system card, 2025a.

OpenAI. OpenAI model spec, September 12, 2025, 2025b.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Aviv Ovadya, Kyle Redman, Luke Thorburn, Quan Ze Chen, Oliver Smith, Flynn Devine, Andrew Konya, Smitha Milli, Manon Revel, Kevin Feng, et al. Position: Democratic AI is possible. the democracy levels framework shows how it might work. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Frank Pasquale. A rule of persons, not machines: The limits of legal automation. *George Washington Law Review*, 87(1):1–55, 2019.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv preprint arXiv:2202.03286*, 2022.

Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. Constitutionmaker: Interactively critiquing large language models by converting feedback into principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 853–868, 2024.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Adriana Placani. Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, 4(3):691–698, 2024.

Eric A Posner and Shivam Saran. Judge AI: assessing large language models in judicial decision-making. 2025.

Vinodkumar Prabhakaran, Margaret Mitchell, Timnit Gebru, and Iason Gabriel. A human rights-based approach to responsible AI. *arXiv preprint arXiv:2210.02667*, 2022.

Dasha Pruss and Jessie Allen. Against AI jurisprudence: Large language models and the false promises of empirical judging. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2055–2066, 2025.

Abhishek Purushothama, Junghyun Min, Brandon Waldon, and Nathan Schneider. Not ready for the bench: LLM legal interpretation is unstable and uncalibrated to human judgments. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, Daniel Preoțiuc-Pietro, and Gerasimos Spanakis (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2025*, pp. 317–317. Association for Computational Linguistics, November 2025.

Ketan Ramakrishnan, Gregory Smith, and Conor Downey. US tort liability for large-scale artificial intelligence damages. RAND, 2024.

John Rawls. *Political Liberalism*. Columbia University Press, 1993.

John Rawls. *A Theory of Justice*. Belknap Press, Harvard University Press, Revised edition, 1999.

Joseph Raz. Legal principles and the limits of law. *Yale Law Journal*, 81:823, 1971.

Joseph Raz. *Practical reason and norms*. Oxford University Press, 1st edition, 1975.

Joseph Raz. *The Authority of Law: Essays on Law and Morality*. Clarendon Press, Oxford University Press, 1979a.

Joseph Raz. The rule of law and its virtue. In *The Authority of Law: Essays on Law and Morality*, pp. 210–229. Clarendon Press, Oxford University Press, 1979b.

Joseph Raz. The law's own virtue. *Oxford Journal of Legal Studies*, 39(1):1–15, 2019.

Joel R. Reidenberg. Lex informatica: The formulation of information policy rules through technology. *Texas Law Review*, 76:553–593, 1998.

Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel J Kochenderfer. BetterBench: Assessing AI benchmarks, uncovering issues, and establishing best practices. *Advances in Neural Information Processing Systems*, 37:21763–21813, 2024.

Mark O Riedl and Deven R Desai. AI agents and the law. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 2189–2198, 2025.

Edwina L Rissland. Artificial intelligence and law: Stepping stones to a model of legal reasoning. *Yale Law Journal*, 99:1957–1981, 1990.

Stuart Russell. *Human Compatible: AI and the Problem of Control*. Viking, 2019.

Greg Sadler and Nathan Sherburn. Legal zero-days: A novel risk vector for advanced AI systems. *arXiv preprint arXiv:2508.10050*, 2025.

Olawale Salaudeen, Anka Reuel, Ahmed Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. *arXiv preprint arXiv:2505.10573*, 2025.

Peter N. Salib. AI outputs are not protected speech. *Washington University Law Review*, 102:83, 2024.

Peter N. Salib and Simon Goldstein. AI rights for human safety. *Virginia Law Review (forthcoming)*, 2025a.

Peter N. Salib and Simon Goldstein. AI rights for economic flourishing. 2025b.

Keenan Samway, Rada Mihalcea, and Zhijing Jin. When do language models endorse limitations on universal human rights principles? In *Workshop on Socially Responsible Language Modelling Research*, 2025.

Nathan Sanders and Bruce Schneier. AI will write complex laws. *Lawfare*, 2025.

Atrisha Sarkar, Andrei Ioan Muresanu, Carter Blair, Aaryam Sharma, Rakshit S. Trivedi, and Gillian K. Hadfield. Normative modules: A generative agent architecture for learning norms that supports multi-agent cooperation. *arXiv preprint arXiv:2405.19328*, 2024.

Antonin Scalia and Bryan A. Garner. *Reading Law: The Interpretation of Legal Texts*. Thomson West, 2012.

Frederick Schauer. Precedent. *Stanford Law Review*, 39(3):571–605, 1987.

Frederick Schauer. *Playing by the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Clarendon Press, Oxford University Press, 1991.

Frederick Schauer. Giving reasons. *Stanford Law Review*, 47:633–659, 1995.

Frederick Schauer. *Thinking like a lawyer: a new introduction to legal reasoning*. Harvard University Press, 2009.

Jérémy Scheurer, Mikita Balesni, and Marius Hobbhahn. Large language models can strategically deceive their users when put under pressure. *arXiv preprint arXiv:2311.07590*, 2024.

Bruce Schneier. The coming AI hackers. Belfer Center for Science and International Affairs, Harvard Kennedy School, April 2021.

Daniel Schwarcz, Sam Manning, Patrick Barry, David R Cleveland, JJ Prescott, and Beverly Rich. AI-powered lawyering: AI reasoning models, retrieval augmented generation, and the future of legal practice. 2025.

Steven L. Schwarcz. Systemic risk. *Georgetown Law Journal*, 97:193–249, 2008.

Elizabeth Seger, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. Democratising AI: Multiple meanings, goals, and methods. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 715–722, 2023.

Amartya Sen. Human rights and the limits of law. *Cardozo Law Review*, 27:2913, 2005.

Richard Sentinella and Cobun Zweifel-Keegan. US state AI governance legislation tracker. IAPP, July 2025.

Scott J. Shapiro. What is the internal point of view? *Fordham Law Review*, 75:1157–1170, 2006.

Scott J. Shapiro. *Legality*. Belknap Press, Harvard University Press, 2011.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, et al. Towards understanding sycophancy in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Abhay Sheshadri, John Hughes, Julian Michael, Alex Mallen, Arun Jose, Fabien Roger, et al. Why do some language models fake alignment while others don't? *arXiv preprint arXiv:2506.18032*, 2025.

Herbert A. Simon. *Administrative Behavior*. Free Press, Simon & Schuster, 4th edition, 1997.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–6, 2022.

Brad Smith. Microsoft announces new Copilot Copyright Commitment for customers, September 2023.

Nathalie A. Smuha. *Algorithmic rule by law: How algorithmic regulation in the public sector erodes the rule of law*. Cambridge University Press, 2024.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15725–15788, 2024.

Lawrence B. Solum. Legal personhood for artificial intelligences. *North Carolina Law Review*, 70:1231–1287, 1992.

Ziming Song. Value-aligned but misguided: a dilemma in AI and AGI decision making. *Synthese*, 206(3):138, 2025.

Taylor Sorensen, Liwei Jiang, Jena D Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, et al. Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19937–19947, 2024a.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. Position: A roadmap to pluralistic alignment. In *International Conference on Machine Learning*, pp. 46280–46302, 2024b.

Tobin South, Subramanya Nagabhushanaradhya, Ayesha Dissanayaka, Sarah Cecchetti, George Fletcher, Victor Lu, Aldo Pietropaolo, Dean H Saxe, Jeff Lombardo, Abhishek Maligehalli Shivalingaiah, et al. Identity management for agentic AI: The new frontier of authorization, authentication, and security for an AI agent world. *arXiv preprint arXiv:2510.25819*, 2025.

Kevin M. Stack. The constitutional foundations of Chenery. *Yale Law Journal*, 116:952–1040, 2007.

Marilyn Strathern. 'Improving ratings': audit in the british university system. *European Review*, 5(3):305–321, 1997.

Jonathan Stray, Alon Halevy, Parisa Assar, Dylan Hadfield-Menell, Craig Boutilier, Amar Ashar, Chloe Bakalar, Lex Beattie, Michael Ekstrand, Claire Leibowicz, Connie Moon Sehat, Sara Johansen, Lianne Kerlin, David Vickrey, Spandana Singh, Sanne Vrijenhoek, Amy Zhang, McKane Andrus, Natali Helberger, Polina Proutskova, Tanushree Mitra, and Nina Vasan. Building human values into recommender systems: An interdisciplinary synthesis. *ACM Transactions on Recommender Systems*, 2(3), June 2024.

Christopher Summerfield, Lisa P Argyle, Michiel Bakker, Teddy Collins, Esin Durmus, Tyna Eloundou, Iason Gabriel, Deep Ganguli, Kobi Hackenburg, Gillian K Hadfield, et al. The impact of advanced AI systems on democracy. *Nature Human Behaviour*, pp. 1–11, 2025.

Cass R Sunstein. On analogical reasoning. *Harvard Law Review*, 106(3):741–791, 1993.

Cass R Sunstein. Incompletely theorized agreements. *Harvard Law Review*, 108(7):1733–1772, 1995.

Cass R. Sunstein. Of artificial intelligence and legal reasoning. *University of Chicago Law School Roundtable*, 8:29–45, 2001.

Cass R Sunstein. Artificial intelligence and the first amendment. *George Washington Law Review*, 92:1207, 2024.

Faiz Surani, Lindsey A Gailmard, Allison Casasola, Varun Magesh, Emily J Robitschek, and Daniel E Ho. What is the law? a system for statutory research (STARA) with large language models. In *20th International Conference on Artificial Intelligence and Law*, 2025.

Richard Susskind. *Expert systems in law: a jurisprudential inquiry.* Clarendon Press, Oxford University Press, 1987.

Roberto Tallarita. AI is testing the limits of corporate governance. *Harvard Business Review*, 2023.

Brian Z Tamanaha. *On the rule of law: History, politics, theory.* Cambridge University Press, 2004.

Michael Henry Tessler, Michiel A. Bakker, Daniel Jarrett, Hannah Sheahan, Martin J. Chadwick, Raphael Koster, Georgina Evans, Lucy Campbell-Gillingham, Tantum Collins, David C. Parkes, Matthew Botvinick, and Christopher Summerfield. AI can help humans find common ground in democratic deliberation. *Science*, 386(6719):2852, October 2024.

Michael Henry Tessler, Georgina Evans, Michiel A Bakker, Iason Gabriel, Sophie Bridgers, Rishub Jain, Raphael Koster, Verena Rieser, Anca Dragan, Matthew Botvinick, et al. Can AI mediation improve democratic deliberation? *arXiv preprint arXiv:2601.05904*, 2026.

Rachel Thomas and David Uminsky. The problem with metrics is a fundamental problem for AI. *arXiv preprint arXiv:2002.08512*, 2020.

Nenad Tomasev, Matija Franklin, Joel Z Leibo, Julian Jacobs, William A Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies. *arXiv preprint arXiv:2509.10147*, 2025.

Alan Turing. Computing machinery and intelligence. *Mind: A Quarterly Review of Psychology and Philosophy*, 59(236):433–460, 1950.

Tom R Tyler. *Why people obey the law.* Princeton University Press, 2nd edition, 2006.

U.S. Supreme Court. United states v. kirby. 74 U.S. (7 Wall.) 482, 1868.

Risto Uuk, Carlos Ignacio Gutierrez, Daniel Guppy, Lode Lauwaert, Atoosa Kasirzadeh, Lucia Velasco, Peter Slattery, and Carina Prunkl. A taxonomy of systemic risks from general-purpose AI. *arXiv preprint arXiv:2412.07780*, 2024.

Julian Velasco. The fundamental rights of the shareholder. *UC Davis Law Review*, 40:407, 2006.

Brandon Waldon, Nathan Schneider, Ethan Wilcox, Amir Zeldes, and Kevin Tobia. Large language models for legal interpretation? don't take their word for it. *Georgetown Law Journal (forthcoming)*, 2025.

Jeremy Waldron. The rule of law. *Stanford Encyclopedia of Philosophy*, 2016.

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024.

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. Position: Evaluating generative AI systems is a social science measurement challenge. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Alexander Wan, Kevin Klyman, Sayash Kapoor, Nestor Maslej, Shayne Longpre, Betty Xiong, Percy Liang, and Rishi Bommasani. The 2025 foundation model transparency index. *arXiv preprint arXiv:2512.10169*, 2025.

Xintong Wang and Michael P Wellman. Market manipulation: An adversarial learning framework for detection and evasion. In *29th International Joint Conference on Artificial Intelligence*, 2020.

Yiding Wang, Yuxuan Chen, Fanxu Meng, Xifan Chen, Xiaolei Yang, and Muhan Zhang. Law in silico: Simulating legal society with LLM-based agents. *arXiv preprint arXiv:2510.24442*, 2025.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

Kevin Wei and Lennart Heim. Designing incident reporting systems for harms from general-purpose AI. *arXiv preprint arXiv:2511.05914*, 2025.

Kevin Wei, Patricia Paskov, Sunishchal Dev, Michael J Byun, Anka Reuel, Xavier Roberts-Gaal, Rachel Calcott, Evie Coxon, and Chinmay Deshpande. Position: Human baselines in model evaluations need rigor and transparency (with recommendations & reporting checklist). In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pp. 214–229, 2022.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, et al. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*, 2023.

Laura Weidinger, Inioluwa Deborah Raji, Hanna Wallach, Margaret Mitchell, Angelina Wang, Olawale Salaudeen, Rishi Bommasani, Deep Ganguli, Sanmi Koyejo, and William Isaac. Toward an evaluation science for generative AI systems. *arXiv preprint arXiv:2503.05336*, 2025.

Gabriel Weil. Tort law as a tool for mitigating catastrophic risk from artificial intelligence. 2024.

Daniel Wilf-Townsend and Kevin Tobia. AI-generated legal texts, May 2025.

Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan. On targeted manipulation and deception when optimizing LLMs for user feedback. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Sophie Williams, Jonas Schuett, and Markus Anderljung. On regulating downstream AI developers. *European Journal of Risk Regulation*, pp. 1–29, 2025b.

Tim Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology Law*, 2:141–175, 2003.

Xinyi Wu, Geng Hong, Pei Chen, Yueyue Chen, Xudong Pan, and Min Yang. Prison: Unmasking the criminal potential of large language models. *arXiv preprint arXiv:2506.16150*, 2025.

Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

Yi Zeng, Yu Yang, Andy Zhou, Jeffrey Ziwei Tan, Yuheng Tu, Yifan Mai, Kevin Klyman, Minzhou Pan, Ruoxi Jia, Dawn Song, Percy Liang, and Bo Li. Air-bench 2024: A safety benchmark based on regulation and policies specified risk categories. In *The Thirteenth International Conference on Learning Representations*, 2025.

Andy K Zhang, Neil Perry, Riya Dulepet, Joey Ji, Celeste Menders, Justin W Lin, Eliot Jones, Gashon Hussein, Samantha Liu, Donovan Julian Jasper, Pura Peetathawatchai, Ari Glenn, Vikram Sivashankar, Daniel Zamoshchin, Leo Glikbarg, Derek Askaryar, Haoxiang Yang, Aolin Zhang, Rishi Alluri, Nathan Tran, Rinnara Sangpisit, Kenny O Oseleononmen, Dan Boneh, Daniel E. Ho, and Percy Liang. Cybench: A framework for evaluating cybersecurity capabilities and risks of language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Lucia Zheng, Neel Guha, Javokhir Arifov, Sarah Zhang, Michal Skreta, Christopher D Manning, Peter Henderson, and Daniel E Ho. A reasoning-focused legal retrieval benchmark. In *Proceedings of the 2025 Symposium on Computer Science and Law*, pp. 169–193, 2025.

Tan Zhi-Xuan, Micah Carroll, Matija Franklin, and Hal Ashton. Beyond preferences in AI alignment. *Philosophical Studies*, 182(7):1813–1863, 2025.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy K Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Antony Kellermann, Jasjeet S Sekhon, Jacob Steinhardt, Sarah Schwettmann, Arvind Narayanan, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices in building rigorous agentic benchmarks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025a.

Yuxuan Zhu, Antony Kellermann, Dylan Bowman, Philip Li, Akul Gupta, Adarsh Danda, Richard Fang, Conner Jensen, Eric Ihli, Jason Benn, Jet Geronimo, Avi Dhir, Sudhit Rao, Kaicheng Yu, Twm Stone, and Daniel Kang. CVE-bench: A benchmark for AI agents' ability to exploit real-world web application vulnerabilities. In *Forty-second International Conference on Machine Learning*, 2025b.

Jonathan Zittrain. *The Future of the Internet—And How to Stop It.* Yale University Press, 2008.

Jonathan Zittrain. We need to control AI agents now, July 2024.