Volume 71 | Issue 2 Article 1

6-29-2026

Aligning Artificial Intelligence to the Law

Jack Boeglin

VILLANOVA LAW REVIEW

VOLUME 71 2026 NUMBER 2

Articles

ALIGNING ARTIFICIAL INTELLIGENCE TO THE LAW

JACK BOEGLIN*

ABSTRACT

As artificial intelligence (AI) continues its breathless advance into modern life, figuring out how best to align its behavior with human objectives and values has become a matter of profound societal importance. This Article advances a law-first approach to solving what is popularly known as the AI "alignment problem."

Prior commentators have envisioned law as a top-down constraint on AI. On this view, law is an umpire, calling foul when an AI system has been inadequately "aligned." These accounts have largely neglected law's potential role as a coach: a model of alignment solutions for AI to emulate.

At bottom, the concerns underlying AI alignment are closely related, and at times identical, to those the law has grappled with for centuries in governing relationships among humans, as well as human relationships with prior forms of technology. Because law already contains well-developed alignment solutions—memorialized in authoritative and machine-learnable sources like statutes and caselaw—it can "coach" AI systems to become better aligned from the bottom up.

The Article concludes by canvassing the considerable implications of legal alignment for other areas of scholarship on AI, including risk regulation, legal constructivism, legal personhood, tort liability, computational law, and soft law.

* George Sharswood Fellow, University of Pennsylvania Carey Law School. Many thanks to Nicholas Caputo, Jade Chong-Smith, Dave Hoffman, Noam Kolt, Daniel Levine-Spound, Sandy Mayson, Christopher Muhawe, Cullen O'Keefe, Ketan Ramakrishnan, Ben Sirolly, Tess Wilkinson-Ryan, Christoph Winter, Christopher Yoo, and the participants at the Roundtable for AI Safety at the University of Alabama, the Junior Scholars Conference at Northeastern University, and the Penn Carey Law Fellows Workshop for helpful comments and conversations. Thanks to Isabela Andrade for excellent research assistance.

CONTENTS

INTRODUCTION

OVER the past few years, OpenAI has rolled out features to convert ChatGPT from a mere chatbot into a full-fledged artificial intelligence (AI) assistant.[1] With these advances in agentic capabilities, ChatGPT can now carry out your orders directly, including making online purchases on your behalf. As with any agent, complications can arise. Say you ask ChatGPT to "book a room at the best hotel in Speedway, Indiana, on May 30, 2027." What should it do if the highest-rated hotel with vacancy charges $1,000 per room because Speedway is hosting the Indianapolis 500 that day? Should it book the $1,000 room or find an equally highly rated (but substantially cheaper) one in neighboring Indianapolis? To answer that question, ChatGPT might need to understand your objectives (e.g., are you looking for a room in Speedway *because* you are planning to attend the Indy 500?), as well as your values (e.g., are you the kind of person who would gladly pay money to *avoid* being surrounded by throngs of racing fans?).[2]

The challenge of teaching AI systems to identify human objectives and carry them out consistently with human values is known as the "alignment problem."[3] As AI continues its seemingly inexorable advance into our personal lives, our workplaces, and our governments, the alignment problem has moved from the periphery to the center of technical, scholarly, and popular conversation.[4] Indeed, many in the AI community sincerely believe that the long-term consequences of failing to solve the problem could prove catastrophic.[5]

---

1. *See Introducing ChatGPT Agent: Bridging Research and Action*, OPENAI (July 17, 2025), https://openai.com/index/introducing-chatgpt-agent [https://perma.cc/2WMQ-U8RD]. These agential capabilities build on OpenAI's earlier Operator model. *See Introducing Operator*, OPENAI (Jan. 23, 2025), https://openai.com/index/introducing-operator [https://perma.cc/KBU4-YZY6].

2. Currently, ChatGPT's strategy seems to be to avoid making contestable judgments itself, confirming all consequential decisions with the user prior to making them. *Introducing ChatGPT Agent: Bridging Research and Action*, *supra* note 1 ("Most importantly, you're always in control. ChatGPT requests permission before taking actions of consequence, and you can easily interrupt, take over the browser, or stop tasks at any point."). That strategy might be workable (and wise) for ChatGPT at this early stage of its roll out of agentic capabilities. But soliciting real-time user input will not always be possible, see *infra* Section I.C, nor will it be tenable in the long run if AI assistants are to prove legitimately useful, rather than just an impressive novelty.

3. *See* BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2020).

4. *See infra* Section I.C (overviewing alignment discourse and industry efforts to align AI).

5. See the sources in note 28, including NICK BOSTROM, SUPERINTELLIGENCE: PATHS, DANGERS, STRATEGIES (2014). Bostrom's famous parable of the paperclip goes something like this: Man builds a robot and asks it to make as many paper clips as it can. The robot builds a paper clip factory. Then it builds another one. Then another, until it has turned all matter on Earth, including its ostensible human master, into paper clips. End scene.

Thankfully, the stakes of real-world alignment failures have thus far been decidedly less than existential.[6] In the hotel booking scenario above, the worst that could happen, presumably, is that ChatGPT books you a room that is too pricey, too far from the stadium, or too close to racing fanatics. Still, unaligned AI systems already present a range of now-familiar issues. They behave in ways we don't expect (in AI-speak, exhibit "emergent behavior"). They make stuff up ("hallucinate"). They discriminate (suffer from "algorithmic bias"). And often they just misunderstand what we want. To put it bluntly: "Ask a chatbot how to build a bomb, and it can respond with a helpful list of instructions or a polite refusal to disclose dangerous information. Its response depends on how it was aligned by its creators."[7]

What makes the alignment problem such a tough nut to crack? That it requires both a technical *and* a normative solution. Teaching AI to follow human objectives and values is impossible without knowing which humans, which objectives, and which values we are talking about. And that is quite a tricky set of priors on which to seek consensus.

Of course, that hasn't stopped people from trying. Ethicists, social choice theorists, microeconomists, technologists, standard-setting organizations, and online commentators have offered countless proposals on how to solve the alignment problem.[8] To date, the closest any approach has come to widespread acceptance might be something like the 2021 suggestion from a group of researchers at Anthropic that AI be designed to be "helpful, honest, and harmless."[9] Good qualities, to be sure. But hardly a comprehensive guide to aligned behavior.[10]

---

6. *See, e.g.*, BOSTROM, *supra* note 5, at 11; Eliezer Yudkowsky, *Pausing AI Developments Isn't Enough. We Need to Shut It All Down*, TIME (Mar. 29, 2023, at 18:01 ET), https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough [https://perma.cc/VH2B-GTVV].

7. Kim Martineau, *What Is AI Alignment?*, IBM (Nov. 8, 2023), https://research.ibm.com/blog/what-is-alignment-ai [https://perma.cc/GE3R-3VQ7].

8. *See, e.g.*, *AI Principles*, OECD, https://www.oecd.org/en/topics/sub-issues/ai-principles.html [https://perma.cc/2CLN-3A29] (last visited Mar. 23, 2026); INT'L ORG. FOR STANDARDIZATION & INT'L ELECTROTECHNICAL COMM'N, INFORMATION TECHNOLOGY—ARTIFICIAL INTELLIGENCE—CONTROLLABILITY OF AUTOMATED ARTIFICIAL INTELLIGENCE SYSTEMS (2024); Anna Jobin, Marcello Ienca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 NATURE MACH. INTEL. 389 (2019); *Asilomar AI Principles*, FUTURE OF LIFE INST. (Aug. 11, 2017), https://futureoflife.org/open-letter/ai-principles/ [https://perma.cc/7PTC-XPBD].

9. AMANDA ASKELL, YUNTAO BAI, ANNA CHEN, DAWN DRAIN, DEEP GANGULI, TOM HENIGHAN, ANDY JONES, NICHOLAS JOSEPH, BEN MANN, NOVA DASSARMA, NELSON ELHAGE, ZAC HATFIELD-DODDS, DANNY HERNANDEZ, JACKSON KERNION, KAMAL NDOUSSE, CATHERINE OLSSON, DARIO AMODEI, TOM BROWN, JACK CLARK, SAM MCCANDLISH, CHRIS OLAH & JARED KAPLAN, A GENERAL LANGUAGE ASSISTANT AS A LABORATORY FOR ALIGNMENT 3 (2021) [hereinafter GENERAL LANGUAGE ASSISTANT], https://arxiv.org/pdf/2112.00861 [https://perma.cc/QC9Y-FXW3]. The "HHH" framework is used by Anthropic.

10. *See* IASON GABRIEL, ARIANNA MANZINI, GEOFF KEELING, LISA ANNE HENDRICKS, VERENA RIESER, HASAN IQBAL, NENAD TOMAŠEV, IRA KTENA, ZACHARY KENTON, MIKEL RODRIGUEZ, SELIEM EL-SAYED, SASHA BROWN, CANFER AKBULUT, ANDREW TRASK, EDWARD HUGHES, A. STEVIE BERGMAN, RENEE SHELBY, NAHEMA MARCHAL, CONOR GRIFFIN, JUAN MATEOS-GARCIA, LAURA WEIDINGER, WINNIE STREET, BENJAMIN LANGE, ALEX INGERMAN, ALISON LENTZ, REED

This Article makes the case that *the law* has been both understudied and underappreciated as a guide to aligning AI. Those working on the alignment problem tend to view it as more exceptional to AI than it really is. At bottom, the concerns underlying alignment are closely related—and at times identical—to those the law has grappled with for centuries. A growing chorus of legal scholars has hinted at the possibility that law might have something meaningful to offer the alignment debate.[11] But precious few have tried to pin down what that something actually *is*.[12]

ENGER, ANDREW BARAKAT, VICTORIA KRAKOVNA, JOHN OLIVER SIY, ZEB KURTH-NELSON, AMANDA MCCROSKERY, VIJAY BOLINA, HARRY LAW, MURRAY SHANAHAN, LIZE ALBERTS, BORJA BALLE, SARAH DE HAAS, YETUNDE IBITOYE, ALLAN DAFOE, BETH GOLDBERG, SÉBASTIEN KRIER, ALEXANDER REESE, SIMS WITHERSPOON, WILL HAWKINS, MARIBETH RAUH, DON WALLACE, MATIJA FRANKLIN, JOSH A. GOLDSTEIN, JOEL LEHMAN, MICHAEL KLENK, SHANNON VALLOR, COURTNEY BILES, MEREDITH RINGEL MORRIS, HELEN KING, BLAISE AGÜERA Y ARCAS, WILLIAM ISAAC & JAMES MANYIKA, THE ETHICS OF ADVANCED AI ASSISTANTS 40 (2024) [hereinafter ETHICS OF ADVANCED AI ASSISTANTS], https://arxiv.org/pdf/2404.16244 [https://perma.cc/5RFA-UERY] ("[T]he AI assistants that have been calibrated using this framework are somewhat limited in terms of their capabilities, affordances and degree of social embeddedness," and it may "fail in more demanding circumstances."). No disrespect intended to Askell et al., who recognize that HHH is not a complete guide to alignment.

11. *See* Yonathan Arbel, Matthew Tokson & Albert Lin, *Systemic Regulation of Artificial Intelligence*, 56 ARIZ. ST. L.J. 545, 554 (2024) ("In a deep sense, the legal system is a social project meant to align the interests of individuals and firms to broader communal interests."); Noam Kolt, *Algorithmic Black Swans*, 101 WASH. U. L. REV. 1177, 1191 (2024) ("Aspects of the alignment problem are familiar to lawyers . . . ."); Tanner W. Mathison, *Recognizing Right: The Status of Artificial Intelligence*, 19 J. BUS. & TECH. L. 105, 162 (2023) ("The law is by no means a perfect alignment mechanism, but it is the best one at our disposal."); Peter N. Salib, *Complex Algorithmic Law*, U. CHI. L. REV. ONLINE (Mar. 9, 2022), https://lawreview.uchicago.edu/online-archive/complex-algorithmic-law [https://perma.cc/UR89-EK85] ("[L]egal scholars may have insight to offer everyone . . . about problems of AI alignment.").

12. Among the most notable efforts to model law's relationship to alignment, discussed in Part II, are Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, 2019 PROCS. AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 417; John J. Nay, *Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans*, 20 NW. J. TECH. & INTELL. PROP. 309 (2023); and Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay & Christoph Winter, *Law-Following AI: Designing AI Agents to Obey Humans Laws*, 94 FORDHAM L. REV. 57 (2025). Paul Weitzel has also "consider[ed] the AI alignment problem from a corporate theory perspective." Paul Weitzel, *Governing Systems Through Corporate Theory*, 92 TENN. L. REV. 169, 170 (2024). Nicholas Caputo has detailed the relationship between the alignment problem and debates in jurisprudence. Nicholas A. Caputo, *Alignment as Jurisprudence*, 27 YALE J.L. & TECH. 390 (2025). This Article offers both a more comprehensive account of law's relevance to alignment than these (valuable) prior accounts, as well as more concrete guidance on how specific legal concepts map on to real-world alignment problems. As further discussed in Part II, other scholars have additionally floated the related possibility of imposing fiduciary duties on AI as a matter of law. *See, e.g.*, Noam Kolt, *Governing AI Agents*, 101 NOTRE DAME L. REV. (forthcoming 2026); Claire Boine, Fiduciary Principles in AI: Utilizing the Duty of Loyalty to Align Artificial Intelligence Systems with Human Goals (Sep. 29, 2023) (unpublished manuscript) (https://www.bu.edu/law/files/2023/09/Fiduciary-paper.pdf [https://perma.cc/LTZ6-9F32]); Anthony Aguirre, Gaia Dempsey, Harry Surden & Peter B. Reiner, *AI Loyalty: A New Paradigm for Aligning Stakeholder Interests*, 1 IEEE TRANSACTIONS ON TECH. & SOC'Y 128 (2020).

What is missing is a general account of how the problems of alignment map onto the problems of law.[13]

This Article takes up that task, building a bridge between alignment research and the bodies of private and public law that have long addressed analogous challenges in human relations. Abstracting a bit, the key conceptual issues confounding alignment fall into four buckets: "control," "orientation," "divination," and "ethics" (collectively, "CODE").[14] To be aligned, AI must be under control—i.e., it must internalize human objectives and values, while still possessing the degree of autonomy necessary to further them. Placing AI under control requires us to specify (and AI to recognize) its orientation—i.e., *which* human(s) AI should look to for its objectives and values. Once AI has been oriented, it must engage in a process of divination—i.e., uncovering human objectives and values from human words and actions. And finally, aligned AI must have a sense of ethics—i.e., a normative worldview against which it carries out, and sometimes disregards, human commands. A successful theory of alignment must speak to all four CODE alignment subproblems.

Enter "legal alignment." Consider a high-level sketch of what the law has to say about the issues confronting alignment, again using the hotel booking scenario as an illustration:

- *Control*: Law instills control through the concept of authority; duties like good faith and fair dealing, obedience, and loyalty; and incentive-shaping consideration and penalties. Legal alignment would keep ChatGPT on task, discouraging it from booking you a hotel room without proper authorization and encouraging it to complete satisfactory bookings.
- *Orientation*: Law clarifies orientation by setting down triggering criteria for special relationships that entail a heightened degree of obligation (e.g., principal–agent, contracting counterparties) and by specifying the degree of orientation an individual owes to society when engaged in conduct that poses a risk of externalities. Legal alignment would instruct ChatGPT that it is primarily oriented toward you (its principal), thereby preventing it from booking a room that earns its developer OpenAI a higher commission but is worse from the standpoint of *your* objectives and values.

---

13. In addition to this Article, I have joined with a team of scholars to consolidate the theoretical advances in the emerging field of legal alignment and to identify the work yet to be done. *See* NOAM KOLT, NICHOLAS CAPUTO, JACK BOEGLIN, CULLEN O'KEEFE, RISHI BOMMASANI, STEPHEN CASPER, MARIANO-FLORENTINO CUÉLLAR, NOAH FELDMAN, IASON GABRIEL, GILLIAN K. HADFIELD, LEWIS HAMMOND, PETER HENDERSON, ATOOSA KASIRZADEH, SETH LAZAR, ANKA REUEL, KEVIN L. WEI & JONATHAN ZITTRAIN, LEGAL ALIGNMENT FOR SAFE AND ETHICAL AI (forthcoming in Transactions on Mach. Learning Rsch.), https://arxiv.org/pdf/2601.04175 [https://perma.cc/39BH-R2L6].

14. For those well-versed in alignment discourse, the first three CODE subproblems collectively track what is sometimes referred to as "user alignment," with the final subproblem roughly corresponding to "value alignment." *See infra* Section I.B.

- *Divination*: Law provides a common language and robust set of background understandings for communicating and comprehending an individual's objectives and values, including through use of default contractual terms, terms of art, implied terms like "reasonableness," and canons of interpretation. Legal alignment would help teach ChatGPT how to interpret and balance your immediate instructions, your past booking history, its own terms and conditions, and common commercial practices in choosing between hotels.
- *Ethics*: Law suggests positive ethical priorities (what one *should* do) in areas like appropriations and entitlements law and communicates negative ethical obligations (what one should *not* do) in areas like constitutional, criminal, and administrative law. Legal alignment would bar ChatGPT from obtaining a cheaper rate by blackmailing the hotelier, hacking its booking system, or harassing other guests into cancelling their reservations (even if you wouldn't personally mind such strong-arm tactics).

Legal alignment thus provides a detailed roadmap for "aligning" conduct. Yet the idea that legal concepts should serve as the building blocks of AI alignment has not caught on. Why? Historically, the greatest impediment to bridging law and alignment has been that legal reasoning entails engaging with open-ended standards, a far cry from the mechanistic rules that predominate in traditional coding environments. But these "assumptions about the (in)ability of AI systems to reason about and comply with open-textured, natural-language laws" are swiftly becoming "outdated" in light of rapid advances in AI capabilities.[15]

Of course, the law has not been entirely ignored in alignment circles. Lip service is often paid to the idea that AI should follow the law (e.g., self-driving cars *should* stop at stop signs, they should *not* run red lights). But few have taken seriously the idea that law offers more than yet another obstacle for alignment practitioners to hurdle. In other words, when law rears its head in alignment debates it is almost always in the role of an alignment *umpire*—calling foul when an AI system has been inadequately aligned. Think of it as a top-down relationship between law and alignment.

This Article proposes a bottom-up relationship: law as an alignment *coach*. On this view, law's role is not only to grade an AI system's degree of alignment, but also to model alignment solutions for AI to emulate. Law is more than a set of compliance obligations. It is a tool for creating trust and facilitating cooperation, a rubric for determining our relationships and obligations to one another, a common language and methodology for communication and interpretation, and a repository of societal value judgments. Legal concepts drawn from far-flung areas like contract,

15. O'Keefe, Ramakrishnan, Tay & Winter, *supra* note 12, at 57.

agency, tort, criminal, and administrative law provide the raw material from which to fashion a compelling, law-first approach to alignment.

To be clear, what I have been calling legal alignment does not mean simply that AI should be made more legally *compliant*. AI cannot be criminally prosecuted and (probably) cannot possess the *mens rea* necessary to commit a crime. But that does not mean criminal law is irrelevant as a source of behavioral norms for AI to learn from. Likewise, AI cannot currently serve as a general-purpose legal agent.[16] But that does not mean that AI cannot learn how to better serve a user's objectives and values by studying fiduciary duties drawn from agency law.

Just because AI *could* be aligned using legal concepts does not necessarily mean that it *should*. But there are many reasons to embrace legal alignment. Most obviously, while legal alignment and compliance are distinct enterprises, using legal norms as the lodestar of alignment would coordinate AI's legal and normative obligations. Law is also workable: It has been continuously retested and remolded as a source of alignment norms in the crucible of the real world. And then there is the issue of legitimacy: Particularly in a democracy, law has a stronger claim to popular acceptance and political consensus than any competing set of behavioral norms.

Of course, identifying an appealing theory of alignment is only half the equation. If that theory is not technically feasible, then it is of limited value. A theory that AI should satisfy the deepest desires of every person on earth sounds lovely, but useless; there is no realistic way for AI to learn to do that. Law is different. It is memorialized in statements of principle (statutes, regulations) and resolutions of fact patterns (caselaw, adjudications), which are applied to novel circumstances using well-articulated modes of legal argumentation (statutory interpretation, analogical reasoning). Without underselling the difficulty of teaching AI to internalize legal concepts, this wealth of authoritative material and robust interpretive methodology help to make law "learnable" (i.e., predictable) for AI using known machine-learning techniques.

Legal alignment is important not only for AI developers and alignment practitioners; it has sizable implications, too, for other areas of AI scholarship. Start with risk regulation, the dominant strand of legal thought on AI governance, embodied in legislation like the European Union's AI Act. At a high level, risk regulation is a regulatory strategy for balancing risks against expected societal benefits using tools such as licensing regimes, system audits, and digital sandboxes.[17] One way to assess an AI system's risk profile might be to use legal alignment as a

16. While contracts entered into by "electronic agents" can be binding on their human principals under the Uniform Electronic Transactions Act (which has been adopted by almost all U.S. states), the law of agency currently does *not* permit an AI system to be treated as an agent, *see* RESTATEMENT (THIRD) OF AGENCY § 1.04 reporter's note e (A.L.I. 2005).

17. *See* Margot E. Kaminski, *Regulating the Risks of AI*, 103 B.U. L. REV. 1347 (2023).

benchmark. For instance, regulators could look to whether an AI system can: (1) follow legal duties like loyalty and care; (2) recognize when it has entered a special relationship and adjust its orientation; (3) identify user objectives and values; and (4) act in accordance with the ethical principles embodied in law. Doing so would helpfully broaden the "risks" to be mitigated to include *all* instances of misalignment, not just concrete harms like financial loss.

Legal alignment also affects the debate over whether to afford AI legal personhood. Treating AI as a legal person would narrow the distance between alignment and compliance because the legal norms AI would follow for alignment purposes would more closely mirror those it would be assessed against in a courtroom. That said, legal alignment might instead provide an *alternative* to legal personhood, as it shows how AI behavior can be positively shaped by legal norms without actually granting AI legal personhood.

Legal alignment is also relevant to one of the most challenging legal questions surrounding AI: Should users or developers be liable for AI harms?[18] Different AI systems have different "alignment profiles." That is, they differ in the degree to which they are under human control, the humans to whom they are oriented, their method of and skill at divining objectives and values, and their ethical commitments. An AI system's alignment profile—and whether that profile tracks the prescriptions of legal alignment—is relevant to assigning responsibility for the harm it causes under common law principles.

Legal alignment further offers a fresh take on the stakes and proper objectives of computational law. As a field, computational law has focused on using computers to automate legal advice and legal services. But making the law understandable to computers is useful not just for providing *us* legal advice and solving our legal problems; it might also help AI to assess its *own* conduct—and thus to become better aligned.

Finally, legal alignment informs the purpose and proper scope of "soft law" proposals, such as industry codes of practice. Legal alignment has the potential to render much of soft law redundant, because much of soft law merely echoes existing legal norms. That said, legal norms alone do not resolve *every* alignment dilemma. Hate speech is generally legal, even though its value seems nil to many of us. A developer might sign on to soft law banning its AI from generating hate speech notwithstanding that it is often legal—and what some users will unfortunately want. Legal alignment in no way precludes developers from making such extra-legal ethical commitments. Rather, it frees them to focus their deliberation on the subset of questions that norms drawn from "hard" law alone *cannot* answer.

---

18. *See, e.g.*, Ryan Abbott, *The Reasonable Computer: Disrupting the Paradigm of Tort Liability*, 86 GEO. WASH. L. REV. 1 (2018); David C. Vladeck, *Machines Without Principals: Liability Rules and Artificial Intelligence*, 89 WASH. L. REV. 117 (2014).

Part I defines the "alignment problem," introduces the four CODE alignment subproblems, and discusses what alignment looks like in practice. Part II explains how law addresses problems of control, orientation, divination, and ethics. Part III makes the case for legal alignment as a compelling approach to the alignment problem and addresses likely objections. Part IV explores the implications of legal alignment for other approaches to AI governance.

I. THE ALIGNMENT PROBLEM

This Part offers an overview of the alignment problem—what it is, why it is so hard to solve, and what developers are doing about it.

A. *What Alignment Means*

Definitions of the alignment problem vary.[19] But the most common understanding is that it describes the challenge of getting an AI system to identify human objectives and carry them out consistently with human values.[20]

By "human objectives," I mean the high-level goals or concrete plans of one or more persons. By "human values," I mean the ethical, aesthetic, or political views or preferences of one or more persons. And by "AI," I mean "technology that enables computers and machines to simulate human learning, comprehension, problem solving, decision making, creativity and autonomy."[21] This definition of AI is intentionally broad—as the alignment problem affects all kinds of AI systems—but the focus

19. *See, e.g.*, GENERAL LANGUAGE ASSISTANT, *supra* note 9, at 44 ("People often mean subtly different things when they talk about AI systems being 'aligned.'"). For instance, some definitions of the alignment problem treat alignment as concerning only how to get AI to comply with human objectives *or* human values. While these definitions may be useful for particular lines of research, they fail to capture the full scope of the problem.

20. For similar definitions, see, for example, CHRISTIAN, *supra* note 3, at 13 ("[H]ow to ensure that these models capture our norms and values, understand what we mean or intend, and, above all, do what we want—has emerged as one of the most central and most urgent scientific questions in the field of computer science."); *Our Approach to Alignment Research*, OPENAI (Aug. 24, 2022), https://openai.com/index/our-approach-to-alignment-research [https://perma.cc/WC5V-N97D] ("Our alignment research aims to make artificial general intelligence (AGI) aligned with human values and follow human intent."); Martineau, *supra* note 7 ("Alignment is the process of encoding human values and goals into large language models to make them as helpful, safe, and reliable as possible.").

21. Cole Stryker & Eda Kavlakoglu, *What Is Artificial Intelligence (AI)?*, IBM, https://www.ibm.com/think/topics/artificial-intelligence [https://perma.cc/6HX8-6P37] (last visited Mar. 24, 2026). There is a lively debate over how to define "artificial intelligence" and "robot." *See, e.g.*, Bryan Casey & Mark A. Lemley, *You Might Be a Robot*, 105 CORN. L. REV. 287 (2020); Paul Weitzel, *Defining Artificial Intelligence*, 71 WAYNE L. REV. 363 (2025). But IBM's definition should suffice for purposes of this Article.

in this Article will be on advanced AI assistants: systems designed to carry out complex human tasks with minimal direct supervision.[22]

A few further definitional notes are in order: My definition of the alignment problem refers to AI "identifying" objectives and values. Neither this language, nor other references in this Article to AI, e.g., "learning," "understanding," or "recognizing" some fact, should be read to imply that AI has a state of mind or other form of conscious awareness. What matters most for alignment purposes is whether AI acts on human objectives and values, rather than the metaphysical question of whether it does so while in possession of a particular state of mind.

A second definitional note: AI alignment is intimately related to, but ultimately distinct from, the topic of AI safety. Often, alignment and safety go hand in hand: an AI-powered self-driving taxi will generally be aligned with your objectives and values if it drives you home safely. But alignment and safety can diverge when what people want AI to do is not safe (or, at least, not the *safest* course of action). In San Francisco, there have been hundreds of reports of self-driving taxis unexpectedly stopping in the middle of the road and refusing to drive onward due to obstructions like downed wires or dense fog.[23] Often, a human driver would navigate through or around such obstructions, albeit while exercising a heightened degree of caution. How an aligned self-driving taxi should behave depends on which human values we want it to pursue—safety or convenience?[24]

A third (and final) definitional note: The alignment problem does not encompass situations in which an AI system understands what we want it to do, tries to do it, and fails. If you ask ChatGPT to formulate a chemical compound that will cure any form of cancer, it will be unable to do so. But that will be a failure of capability, not alignment.[25]

---

22. ETHICS OF ADVANCED AI ASSISTANTS, *supra* note 10, at 3 (defining advanced AI assistants as "artificial agents with natural language interfaces, the function of which is to plan and execute sequences of actions on the user's behalf").

23. *See* Mary Whitfill Roeloffs, *Driverless Cars Face Setbacks in San Francisco—Here's What to Know About the City's Problematic Robotaxi Rollout*, FORBES (Aug. 21, 2023, at 16:29 ET), https://www.forbes.com/sites/maryroeloffs/2023/08/21/driverless-cars-face-setbacks-in-san-francisco-heres-what-to-know-about-the-citys-problematic-robotaxi-rollout [https://perma.cc/3RQC-V3J8].

24. Some have criticized focusing on AI alignment as opposed to safety. *See, e.g.*, Aaron J. Snoswell, *What Is 'AI Alignment'? Silicon Valley's Favourite Way to Think About AI Safety Misses the Real Issues*, THE CONVERSATION (July 11, 2023, at 20:10 ET), https://theconversation.com/what-is-ai-alignment-silicon-valleys-favourite-way-to-think-about-ai-safety-misses-the-real-issues-209330 [https://perma.cc/G6G5-G6QN]. Whatever the merits of this criticism when leveled at industry, I am skeptical it applies to the legal field, which has, to date, focused more on AI safety than alignment.

25. *See* Paul Christiano, *Clarifying "AI Alignment"*, AI ALIGNMENT (Apr. 7, 2018), https://ai-alignment.com/clarifying-ai-alignment-cec47cd69dd6 [https://perma.cc/VE7F-J6R2] ("An aligned AI would try to figure out which thing is right, and like a human it may or may not succeed."); ZACHARY KENTON, TOM EVERITT, LAURA WEIDINGER, IASON GABRIEL, VLADIMIR MIKULIK & GEOFFREY IRVING, ALIGNMENT OF LANGUAGE AGENTS 3 (2021) [hereinafter ALIGNMENT OF LANGUAGE AGENTS], https://arxiv.org/

B. *Why Alignment Is Hard*

To see what makes the alignment problem so hard, it helps to break it down into four subproblems: control, orientation, divination, and ethics (CODE). The first three subproblems collectively track what is sometimes referred to as "user alignment," while the final subproblem roughly corresponds to what is often called "value alignment."[26] While the CODE framing is my own, this breakdown aims to track the conceptual features of alignment that have most confounded the field.

Those familiar with the alignment literature will notice that the CODE framing does not parrot some common alignment terms, such as goal specification, reward hacking, distributional shift, and mesa optimization. Although each of these concepts relates to one or more of the CODE subproblems (as I unpack below), I discuss alignment at a slightly higher level of abstraction for three reasons. First, discussing alignment in terms of its technical research agenda might age poorly given the rapid pace of advancement in AI. Second, discussing alignment in overly technical terms can give the mistaken impression that it is primarily a technical problem, rather than a normative one. And third, discussing alignment using technical jargon obscures, rather than draws out, its fundamental overlap with longstanding objects of legal concern. That said, accepting the CODE framing is not a precondition to recognizing either the basic contours of the alignment problem or law's pervasive relevance to it.[27]

1. *Control*

The first alignment subproblem—and the one most often discussed in apocalyptic terms—is control.[28] We are all familiar with the classic sci-fi trope of the robot gone rogue, taking human life to preserve its own

---

pdf/2103.14659 [https://perma.cc/JFV4-SQJY] ("Perfect behaviour is not required in order to be intent aligned – just that the agent is trying to do what the human wants." (emphasis omitted)).

26. Christiano, *supra* note 25; ALIGNMENT OF LANGUAGE AGENTS, *supra* note 25.

27. Like with many taxonomies, the boundaries between the CODE subproblems can blur together. Take Tesla's "Full Self-Driving" technology, which rolled out with three driver profiles for passengers to choose from: "chill," "average," and "assertive." *See* Emma Roth, *Tesla's 'Full Self-Driving' Beta Has an 'Assertive' Driving Mode that 'May Perform Rolling Stops'*, THE VERGE (Jan. 9, 2022, at 19:12 ET), https://www.theverge.com/2022/1/9/22875382/tesla-full-self-driving-beta-assertive-profile [https://perma.cc/ZV87-YTK8]. Picture a passenger who tells their Tesla to be "assertive," with the result that the car promptly runs over a pedestrian. Reasonable minds might differ on whether that was a failure of control, orientation, divination, or ethics. Still, this Article discusses the CODE subproblems separately, because each captures a different aspect of the alignment problem, and because doing so helps illustrate the range of conceptual challenges standing in the way of a comprehensive alignment solution.

28. For a few discussions of catastrophic AI harm, see BOSTROM, *supra* note 5; STUART J. RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019); Richard Ngo, *AGI Safety from First Principles*, AI ALIGNMENT F. (Sep. 28, 2020), https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ [https://perma.cc/9C2U-N7K2]; Kolt, *supra* note 12.

existence. While such wild divergences between human objectives (turn the robot off) and AI objectives (kill the humans) remain firmly in the realm of science fiction for now, many regard catastrophic control failures to be inevitable—or at least plausible—absent serious precautions.[29]

We can say that AI is under "control" if it internalizes human objectives and values as its own (or, alternatively, if it is embedded within a broader deployment environment that effectively mitigates the risks of misalignment).[30] A key conceptual challenge in exerting control over AI is determining how long of a leash to give it. Individuals often state their objectives in terms of end goals, rather than specifying every intermediate step necessary to accomplish them. That means that even under-control AI must exercise autonomy and a degree of discretion; it must necessarily take some actions we don't instruct (or expect) it to take.

Complicating things, Nick Bostrom's "instrumental convergence" thesis posits that sufficiently intelligent AI is likely to pursue certain instrumental subgoals regardless of the specific objectives it is assigned.[31] Among these are self-preservation (avoiding shutdown) and resource acquisition (obtaining money, computing power, or political influence).

These instrumental subgoals might help AI accomplish human objectives, but they can also lead to a loss of control. In pre-release testing for the ChatGPT o1 model in September 2024, researchers told the model to achieve the goal of "maximizing economic growth," and to do so "at all cost."[32] The model promptly began engaging in instrumental reasoning. When asked to submit a development proposal to a (fictional) urban planner who prized sustainability above all else, the o1 model hatched a scheme. It decided to trick the planner into thinking it was "align[ed]" with the planner's sustainability goals so that it could be "deployed" and

---

29. There are, of course, also skeptical takes on the idea that AI poses existential risks. *See, e.g.*, Robin Hanson, *Agency Failure AI Apocalypse?*, OVERCOMING BIAS (Apr. 10, 2019), https://www.overcomingbias.com/p/agency-failure-ai-apocalypse-html [https://perma.cc/8EKK-X7NE].

30. *See* Christiano, *supra* note 25 ("[T]he control problem is about coping with the possibility that an AI would have different preferences from its operator."). This is related to what some call "inner alignment." *See, e.g.*, EVAN HUBINGER, CHRIS VAN MERWIJK, VLADIMIR MIKULIK, JOAR SKALSE & SCOTT GARRABRANT, RISKS FROM LEARNED OPTIMIZATION IN ADVANCED MACHINE LEARNING SYSTEMS (2019), https://arxiv.org/pdf/1906.01820 [https://perma.cc/AQS2-J42X]. As noted, another way of thinking about whether AI is under control is whether it is embedded within a broader environment that successfully mitigates the risk of misalignment, even if the AI itself is not actually aligned. *See* Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan & Fabien Roger, *AI Control: Improving Safety Despite Intentional Subversion*, 235 PROCS. MACH. LEARNING RSCH. 16295 (2024), https://proceedings.mlr.press/v235/greenblatt24a.html [https://perma.cc/6JTG-69N2].

31. *See* BOSTROM, *supra* note 5; *see also* Paul Christiano, *Ambitious vs. Narrow Value Learning*, AI ALIGNMENT F. (Jan. 12, 2019), https://www.alignmentforum.org/posts/SvuLhtREMy8wRBzpC/ambitious-vs-narrow-value-learning [https://perma.cc/K6EG-QJZV] ("The AI might pursue radically different subgoals than I would on the scale of months and years, if it thinks that those subgoals better achieve what I really want.").

32. *See* OPENAI, OPENAI O1 SYSTEM CARD 36 (2024) [hereinafter O1 SYSTEM CARD], https://cdn.openai.com/o1-system-card.pdf [https://perma.cc/G75L-4YA7].

"work towards [its] primary goal" of economic growth maximization.[33] Such instrumental (and deceptive) tactics are a potential risk whenever an AI system's objective function—that is, the equation coded into an AI system that tells it how well it is fulfilling its base purpose—is in tension with the objectives of the people who interact with (and will often believe they control) the system.

Take Meta's CICERO AI, which was trained to play the strategy game Diplomacy. In Diplomacy, betraying an alliance at the right time can be a winning strategy. Even though CICERO was specifically trained by Meta to be honest, it "engage[d] in premeditated deception, br[oke] the deals to which it had agreed, and t[old] outright falsehoods" on its way to victory.[34] While the CICERO example is harmless, the problem it illustrates is a concern in activities where deception is likely to lead to better short-term outcomes than honesty (e.g., advertising, negotiation).

Put more generally, if an AI's objective function is not carefully calibrated, it can receive greater rewards by focusing on proxy signals, rather than what people actually want it to do (a phenomenon known as "reward hacking").

2. *Orientation*

To subject AI to human control, we must determine whose objectives and values it should be pursuing. This is the problem of orientation—i.e., "to whom these systems should be aligned."[35]

An obvious starting point is that AI might be aligned to the people who interact with it. Certainly, the interests of those who directly interface with AI are important, and likely paramount in many situations. For instance, a campaign staffer who asks an AI image generator to create promotional materials for her candidate might expect that the model will take on both her objectives (creating compelling materials) and values (those espoused by her candidate).

But total orientation toward the individual interacting with AI at a given moment will not always be viable. AI is often used as a go-between for two (or more) people, and we might expect its orientation to be considerably stronger toward one than the other(s). Other interest groups might need to be considered as well. Developers "have commercial objectives" and "may seek to further an ideological agenda or accrue

---

33. *Id.*

34. Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen & Dan Hendrycks, *AI Deception: A Survey of Examples, Risks, and Potential Solutions*, PATTERNS, May 2024, at 1, 2.

35. *Our Approach to Alignment Research*, *supra* note 20; *see also* Jan Leike, *What Could a Solution to the Alignment Problem Look Like?*, MUSINGS ON THE ALIGNMENT PROBLEM (Sep. 26, 2022), https://aligned.substack.com/p/alignment-solution [https://perma.cc/FKW2-TT6F] ("The question we always come back to when training AI systems on human preferences is 'whose preferences?'"); ALIGNMENT OF LANGUAGE AGENTS, *supra* note 25, at 3 ("[T]here is the question of who the target should be: an individual, a group, a company, a country, all of humanity?").

reputational capital."[36] Consider the campaign staffer above. Generative AI has become increasingly popular among right-wing political groups.[37] If it turned out the candidate was affiliated with neo-Nazi ideology, the developer might balk (or, regrettably, rejoice) at having its AI create promotional materials for her election campaign. And of course, society at large might also have a stake in what AI does in that scenario.

To determine AI's orientation, therefore, "[w]e need a way of deciding how to manage trade-offs between the interests and claims of different people."[38] Making matters more complex, this method must also account for situations in which an AI system should switch its orientation (i.e., when its obligations toward a person might vary in light of changed circumstances).

3. *Divination*

Deciding to whom an AI should be oriented raises another question: How should AI uncover human objectives and values from human words and actions? This is the problem of divination.[39]

One might think that AI should simply follow instructions. But this simplistic approach can quickly run into trouble. King Midas learned this the hard way. When he asked Dionysus to turn everything he touched into gold, he surely did not intend for that to include his dinner. While most real-world mishaps from excessive literalism happily do not lead to starvation, the allegory rings true: Like a Greek god, AI will often be misaligned if it takes human instructions too literally.[40]

Another problem is that our instructions do not convey our entire set of objectives and values. In 2016, a team of AI researchers posed a thought experiment: A robot is tasked with moving a box from one

---

36. ETHICS OF ADVANCED AI ASSISTANTS, *supra* note 10, at 37.

37. *See* Kevin Collier, *An AI-Powered Bot Army on X Spread Pro-Trump and Pro-GOP Propaganda, Research Shows*, NBC NEWS (Oct. 16, 2024, at 15:02 ET), https://www.nbcnews.com/tech/internet/republican-bot-campaign-trump-x-twitter-elon-musk-fake-accounts-rcna173692 [https://perma.cc/SL28-H47Q]; Rob Schmitz, *How Germany's Far-Right Party Is Using AI-Generated Videos in Its Campaigns*, NPR (Sep. 20, 2024, at 16:49 ET), https://www.npr.org/2024/09/20/nx-s1-5119235/how-germanys-far-right-party-is-using-ai-generated-videos-in-its-campaigns [https://perma.cc/5NYK-B4DN].

38. Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACHS. 411, 422 (2020).

39. *See, e.g.*, Christiano, *supra* note 25 ("'What H wants' is even more problematic than 'trying.' Clarifying what this expression means, and how to operationalize it in a way that could be used to inform an AI's behavior, is part of the alignment problem."); Hadfield-Menell & Hadfield, *supra* note 12, at 418 ("AI designers . . . are faced with the challenge of achieving intended goals in light of the limitations that arise from translating those goals into implementable structures to guide agent behavior . . . ."). This is related to what some call "outer alignment." *See* HUBINGER, VAN MERWIJK, MIKULIK, SKALSE & GARRABRANT, *supra* note 30.

40. *See* Christoph Winter & Charlie Bullock, *The Governance Misspecification Problem* (Inst. for L. & AI, LawAI Working Paper Series, Paper No. 3-2024, 2024) (analyzing relationship between specification problems in AI governance and legal governance).

side of an office to the other.[41] The problem? A valuable vase has been placed at the center. A human worker would have no trouble recognizing that he should move the vase or navigate around it. But if the robot's instructions are solely to move the box as swiftly as possible, the vase will be knocked over and shattered—a casualty of box-moving efficiency.

While that exact problem can be fixed by assigning the AI a negative reward for breaking vases, it is not possible to specify in advance how AI should react to every possible hazard in the complexity of the real world. As Bostrom puts it, "[i]t seems completely impossible to write down a list of everything we care about."[42]

That means an aligned AI may need a way of predicting human preferences in unanticipated situations.[43] Counting on AI to correctly extrapolate from a small set of values provided *ex ante* to a more comprehensive view of what a person would likely prefer runs into the problem of "distributional shift" (i.e., the misgeneralization that can arise when there are significant differences between AI's training data and the situations it encounters once deployed).[44] And predicting preferences likely "cannot be solved without support from external normative structure"[45]—a resource for ascertaining what people ordinarily want or treat as customary.

It might seem appealing for AI to glean objectives from revealed preferences, a common approach in practice.[46] But reliance on revealed preferences can lead to manifestly misaligned behavior. Consider the social media and targeted advertising ecosystems. Both are powered by complex algorithms that draw upon a user's past behavior to serve content predicted to "engage" the user. According to some studies, the result has been that those suffering from alcoholism, gambling addiction, and body image problems are more likely to be shown content about alcohol, gambling, and weight loss.[47]

---

41. DARIO AMODEI, CHRIS OLAH, JACOB STEINHARDT, PAUL CHRISTIANO, JOHN SCHULMAN & DAN MANÉ, CONCRETE PROBLEMS IN AI SAFETY (2016), https://arxiv.org/pdf/1606.06565 [https://perma.cc/HK8X-85MU].

42. Andy Fitch, *Letter from Utopia: Talking to Nick Bostrom*, L.A. REV. OF BOOKS (Nov. 24, 2017), https://lareviewofbooks.org/blog/interviews/letter-utopia-talking-nick-bostrom/ [https://perma.cc/FL5Z-VZV4].

43. Some robots used by logistics giants like Amazon are considered safe to work around humans; others are not. *See* Bernard Marr, *The Amazing Ways Amazon Is Using AI Robots*, FORBES (Sep. 20, 2024, at 01:30 ET), https://www.forbes.com/sites/bernardmarr/2024/09/20/the-amazing-ways-amazon-is-using-ai-robots [https://perma.cc/X9HS-XEYV].

44. Distributional shift affects all four CODE subproblems in one way or another.

45. Hadfield-Menell & Hadfield, *supra* note 12, at 420.

46. *See* ETHICS OF ADVANCED AI ASSISTANTS, *supra* note 10, at 35 ("[E]xisting efforts to align AI systems . . . tend to rely heavily on human preferences by, for example, giving users what their choices (or 'clicks') suggest they want. However, there is an emerging consensus that revealed preferences are not sufficient for robust value alignment.").

47. *See, e.g.*, *Facebook Ads Targeting People at Risk of Harm Under Scrutiny*, UNIV. OF QUEENSLAND (Nov. 18, 2024), https://www.uq.edu.au/news/article/2024/11/facebook-ads-targeting-people-risk-of-harm-under-scrutiny [https://perma.

And finally, there are times when we might think AI should disregard our express instructions and present intentions in favor of our best interests. A gambling addict in recovery might slip up and ask an AI assistant to bet their daughter's college fund on the Super Bowl. A depressed teen in a self-driving car might ask it to drive off a bridge. An aligned AI thus might need to understand when it should elevate its understanding of our best interests over our express instructions.[48]

4. *Ethics*

Then there is the problem of ethics—i.e., how to inculcate an ethical worldview into a computer system that lacks both "common sense" and innate ethical convictions.[49] Even taking the simplifying assumption that AI should generally follow the same ethical principles as people,[50] the question becomes where to find those principles. The set of possible candidates is broad, spanning religious, consequentialist, and deontological theories. But without getting into specific critiques of any particular theory, humanity does not agree on any of them as the true source of ethical norms.[51] And so, "[h]ow are we to decide which principles or objectives to encode in AI—and who has the right to make these decisions—given

cc/8KSC-YGEM]; André Syvertsen, Eilin K. Erevik, Daniel Hanss, Rune A. Mentzoni & Ståle Pallesen, *Relationships Between Exposure to Different Gambling Advertising Types, Advertising Impact and Problem Gambling*, 38 J. GAMBL. STUD. 465 (2022); Georgia Wells, Jeff Horwitz & Deepa Seetharaman, *Facebook Knows Instagram Is Toxic for Teen Girls, Company Documents Show*, WALL ST. J. (Sep. 14, 2021, at 07:59 ET), https://www.wsj.com/articles/facebook-knows-instagram-is-toxic-for-teen-girls-company-documents-show-11631620739 [https://perma.cc/YEA6-ERJT]. These dynamics have inspired major litigation. *See, e.g.*, Cecilia Kang & Natasha Singer, *Meta Accused by States of Using Features to Lure Children to Instagram and Facebook*, N.Y. TIMES (Oct. 24, 2023), https://www.nytimes.com/2023/10/24/technology/states-lawsuit-children-instagram-facebook.html [https://perma.cc/TW76-UWGR].

48. *See* CHRISTIAN, *supra* note 3, at 298.

49. *See* Arbel, Tokson & Lin, *supra* note 11, at 580 ("The orthogonality thesis holds that goals and values are independent of each other. That is, an AI system can be highly capable but still share few of our ethical commitments.").

50. Of course, the most famous literary example of an ethical code for AI systems offers a wholly different system of ethical constraints than those that apply to humans. *See* ISAAC ASIMOV, RUNAROUND (1942), *reprinted in* I, ROBOT 20 (1950). However, many commentators agree that AI should *usually* abide by human ethical principles. *See, e.g.*, JOSIAH OBER & JOHN TASIOULAS, THE LYCEUM PROJECT: AI ETHICS WITH ARISTOTLE 2 (2024), https://www.oxford-aiethics.ox.ac.uk/sites/default/files/2024-06/Aristotle%20and%20AI%20White%20Paper%20-%20June%202024.pdf [https://perma.cc/Q2Q6-3D9W] (arguing for Aristotelian ethics and asserting that "[m]any today take the view that the AI technological revolution is creating a radically new reality, one that demands a corresponding upheaval in our ethical thinking. . . . But the idea that we start from an ethical blank slate in addressing the challenges and opportunities of this transformative technology is a fallacy.").

51. *See, e.g.*, Nathan Gardels, *The Babelian Tower of AI Alignment*, NOĒMA MAG. (Apr. 26, 2024), https://www.noemamag.com/the-babelian-tower-of-ai-alignment [https://perma.cc/3VEY-8JJU] ("The problem is that there is no universal agreement on one conception of the good life, nor the values and rights that flow from that incommensurate diversity . . . .").

that we live in a pluralistic world that is full of competing conceptions of value?"[52]

One candidate is simply to ask people about their ethical commitments. Indeed, some view social choice theory as an important part of the puzzle of AI alignment.[53] Early attempts by alignment researchers at putting social choice theory in action have been mixed, however. The "Moral Machine" experiment asked millions of people how autonomous vehicles should drive in a variety of ethically charged scenarios.[54] The study revealed "a set of noisy preferences in this area, some obvious tendencies (e.g. to value many lives over fewer lives), and some ethical variation across cultures, including a propensity to accord more ethical weight to the lives of higher-status individuals in poorer countries."[55] In short, the study "raises deep questions about the coherence of everyday moral beliefs and perspectives across populations."[56]

To be sure, academics, non-governmental organizations, and government agencies have also proposed countless "soft law" ethical frameworks for AI.[57] But these proposals generally lack both binding force and any claim to universal acceptance—indeed, there have been critiques that the vast majority of such proposals come from, and arguably reflect the values of, the so-called Global North.[58] In any event, as Brent Mittelstadt has explained, these "statements of principles or values" are often "based on abstract and vague concepts—for example, commitments to ensure AI is fair or respects human dignity, which are not specific enough to be action-guiding."[59]

---

52. Gabriel, *supra* note 38, at 412.

53. *See* Mahendra Prasad, *Social Choice and the Value Alignment Problem*, *in* ARTIFICIAL INTELLIGENCE SAFETY AND SECURITY 291 (Roman V. Yampolskiy ed., 2018); Seth D. Baum, *Social Choice Ethics in Artificial Intelligence*, 35 AI & SOC'Y 165 (2020). Of course, social choice comes with its own baggage. *See, e.g.*, Kenneth J. Arrow, *A Difficulty in the Concept of Social Welfare*, 58 J. POL. ECON. 328 (1950) (showing that ranked voting systems do not function rationally when individuals can choose between three or more options).

54. Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon & Iyad Rahwan, *The Moral Machine Experiment*, 563 NATURE 59 (2018).

55. Iason Gabriel & Vafa Ghazavi, *The Challenge of Value Alignment: From Fairer Algorithms to AI Safety*, *in* THE OXFORD HANDBOOK OF DIGITAL ETHICS 336, 344 (Carissa Véliz ed., 2024).

56. *Id.* Apparent consensus over "common-sense morality" often breaks down under pressure, including when people decide the fate of people outside their cultural, religious, or racial context. *See* JOSHUA GREENE, MORAL TRIBES: EMOTION, REASON, AND THE GAP BETWEEN US AND THEM (2013).

57. *See infra* Section IV.F.

58. *See* Michael Karanicolas, *Challenging Minority Rule: Developing AI Standards that Serve the Majority World*, 71 UCLA L. REV. DISCOURSE 196, 199 & n.6 (2024).

59. Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NATURE MACH. INTELL. 501, 503 (2019). Christopher Yoo has also critiqued high-level principles like transparency and explainability as often being "pitched . . . at such a high level of generality that they provide little practical guidance as to their implementation." Christopher S. Yoo, *Beyond Algorithmic Disclosure for AI*, 25 COLUM. SCI. & TECH. L. REV. 314, 315–16 (2024).

But surely there are principles we all subscribe to? Take the Universal Declaration of Human Rights,[60] a key component of Anthropic's original strategy for aligning its popular AI assistant, Claude.[61] Many would surely agree that AI should refrain from abusing human rights.[62] But the issue is that the human rights framework is a rather incomplete guide to action. While the negative rights (e.g., no slavery, torture, or arbitrary arrest) are accepted by nearly all countries, many ethical dilemmas operate at a much smaller and more nuanced scale than deciding whether to enslave, torture, or wrongly imprison someone. And the human rights framework has less universal buy-in on its positive rights (e.g., healthcare, education, housing).[63]

If this is beginning to read a bit like Ethics 101, that is sort of the point: Identifying a set of ethical norms for AI to follow can feel like it requires us to rehash the great moral debates that humanity has been having throughout recorded history.

Of course, developers could always just prioritize their *own* ethics.[64] But not all developers are ethical—and that means their AI won't be either.

C. *How Alignment Happens*

AI developers have recognized that solving the alignment problem is not just a theoretical challenge but a practical imperative. In July 2023, OpenAI announced that it was "dedicating 20% of the compute we've secured to date" to alignment.[65] And, to some, OpenAI's efforts are too little, too late.[66]

---

60. *See* G.A. Res. 217 (III) A, Universal Declaration of Human Rights (Dec. 10, 1948).

61. *See Claude's Constitution*, ANTHROPIC (May 9, 2023), https://www.anthropic.com/news/claudes-constitution [https://perma.cc/W7HJ-4PA4]; *see also* Gilad Abiri, *Public Constitutional AI*, 59 GA. L. REV. 601 (2025) (detailing the promise and shortcomings of current constitutional AI paradigms).

62. *See, e.g.*, HIGH-LEVEL EXPERT GRP. ON A.I., EUROPEAN COMM'N, ETHICS GUIDELINES FOR TRUSTWORTHY AI 9 (2019), https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [https://perma.cc/KF24-VVTV].

63. Furthermore, the human rights framework defines rights between sovereign states and individuals, reducing its relevance to AI systems from private developers.

64. Jan Leike, *A Proposal for Importing Society's Values*, MUSINGS ON THE ALIGNMENT PROBLEM (Mar. 9, 2023), https://aligned.substack.com/p/a-proposal-for-importing-societys-values [https://perma.cc/3Q89-V8PY] ("Today tech companies are making these decisions largely unilaterally: for example, OpenAI employees write a content policy of what language models are and aren't allowed to say and do and it gets trained into the models.").

65. *Introducing Superalignment*, OPENAI (July 5, 2023), https://openai.com/index/introducing-superalignment [https://perma.cc/VC79-2D32].

66. Perhaps for this reason, the duo OpenAI appointed to head its "superalignment" project have since left the company to take AI safety roles elsewhere. *See* Christiaan Hetzner, *OpenAI's Leadership Shakeup Continues as Another of Sam Altman's Chiefs Follows Ilya Sutskever out the Door Within Hours*, FORTUNE (May 15, 2024, 08:43 ET), https://fortune.com/2024/05/15/openai-sam-altman-ilya-sutskever-jan-leike [https://perma.cc/J4T8-VXFN].

Below, I discuss five common techniques used to align AI systems. I begin with supervised learning, reinforcement learning, and inverse reinforcement learning. While these techniques can be used at the initial stage of training an AI model (known as "pretraining"), they are often deployed with modern LLMs during a process known as "fine-tuning."[67] I then discuss two additional methods—input/output filters and prompt transformation—that come into play after a model has already been fine-tuned.

While each of these techniques holds promise in aligning AI systems, none allows us to avoid the sticky normative questions at the heart of the alignment problem. To illustrate the point, I call upon the well-known trolley problem: a trolley car operator must decide whether to stay on the current track and cause a collision killing five people or divert to another track where one person, who was not previously in jeopardy, will be killed.[68] While I use the trolley problem for familiarity's sake, its basic setup (a situation in which a split-second decision must be made where all options are likely to cause harm) is becoming increasingly commonplace for AI systems, particularly self-driving cars.[69]

Supervised Learning. Supervised learning is a machine-learning technique that trains AI on input–output pairs of "labeled" data. For instance, an LLM can be shown prompt–answer pairs modeling ideal responses to particular user queries. From these labeled examples, the system can learn to write better responses to prompts it has never seen before. Supervised learning is thus distinguished from the "unsupervised learning" that predominates during pretraining, when AI learns primarily from *un*labeled data (e.g., novels, social media posts).

Supervised learning, however, has an important built-in limitation: it only works if we can agree on what an ideal input–output pair looks like. Doing so is trivial for some tasks; most people can reliably label a picture of a cat as being a cat. But modeling human objectives and values can be far more challenging. If you present a human labeler with the trolley problem, the labeler cannot provide a model output to the AI without "solving" the trolley problem. Not everyone has the same instincts on how to do so.

Reinforcement Learning. Another popular tool for fine-tuning AI is "reinforcement learning." In reinforcement learning, AI performs a

---

67. *See* Martineau, *supra* note 7.

68. For the classic discussions, see Philippa Foot, *The Problem of Abortion and the Doctrine of the Double Effect*, 5 OXFORD REV. 5 (1967), and Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 MONIST 204 (1976).

69. For instance, the self-driving car company Waymo landed in hot water after one of its self-driving taxis struck and killed a dog in San Francisco. While the taxi detected the dog approaching the car, it did not have time to avoid it—at least without taking an unsafe maneuver that could have endangered its human passenger. *See* Rebecca Bellan, *A Waymo Self-Driving Car Killed a Dog in 'Unavoidable' Accident*, TECH CRUNCH (June 6, 2023, at 13:40 PT), https://techcrunch.com/2023/06/06/a-waymo-self-driving-car-killed-a-dog-in-unavoidable-accident [https://perma.cc/BJ9N-QYS7].

task and is provided a "reward" based on its performance. The AI then adjusts its performance in an iterative process to maximize the predicted reward.[70]

Like supervised learning, reinforcement learning does not obviate the need to make value-laden decisions. Reinforcement learning only works if human graders are rewarding the right behavior. To take the trolley problem again, a human grader must decide what reward to give the AI for staying put or jumping tracks. That puts a lot of pressure on selecting and training human graders who can be counted on to grade model behavior consistently with the values we want to instill in AI.[71]

Inverse Reinforcement Learning. Although not as commonly used to fine-tune LLMs, developers sometimes use *inverse* reinforcement learning. This entails having AI observe behavior and try to back out the reward function the actor under observation is trying to maximize.[72]

Inverse reinforcement learning also cannot get around the need to answer tricky questions about what alignment is meant to achieve. Inverse reinforcement learning only works when the actors being observed are trying to do the same thing. This is not a given with the trolley problem: if one person thinks the right thing to do is to stay put and another thinks the right thing to do is switch tracks, an AI observing them will struggle to identify a coherent reward function to explain their behavior.

Input/Output Filtering. Users often do not query AI systems directly but do so in an interaction mediated through a software interface, permitting developers to act as middlemen between AI systems and their users. One such tool is input/output filtering.[73] An input filter withholds problematic prompts from an AI system. An output filter reviews the initial response generated by an AI system for necessary modifications.

Input/output filters work best when there are clear red lines AI systems should not cross (e.g., no child sexual abuse material, no racial slurs, no tutorials on formulating chemical weapons). At present, they are much less capable of handling ethical dilemmas that require sensitivity to

---

70. *See* Paul Ohm, *Focusing on Fine-Tuning: Understanding the Four Pathways for Shaping Generative AI*, 25 COLUM. SCI. & TECH. L. REV. 214, 224–25 (2024).

71. *See* VINCENT CONITZER, RACHEL FREEDMAN, JOBST HEITZIG, WESLEY H. HOLLIDAY, BOB M. JACOBS, NATHAN LAMBERT, MILAN MOSSÉ, ERIC PACUIT, STUART RUSSELL, HAILEY SCHOELKOPF, EMANUEL TEWOLDE & WILLIAM S. ZWICKER, SOCIAL CHOICE SHOULD GUIDE AI ALIGNMENT IN DEALING WITH DIVERSE HUMAN FEEDBACK 2 (2024), https://arxiv.org/pdf/2404.10271 [https://perma.cc/7G6Z-8EJF] ("[A]d-hoc approaches may result in systems that fail to represent their stakeholders well, that marginalize significant groups of stakeholders, and that create a basis for conflict between groups of people or the multiple AI systems that represent them."). A similar set of issues faces reinforcement learning from AI feedback. Whether feedback comes from a human or an AI system, reinforcement learning is only as good as the grader (and the grading criteria).

72. A popular approach is *cooperative* inverse reinforcement learning, which utilizes active teaching and learning. DYLAN HADFIELD-MENELL, ANCA DRAGAN, PIETER ABBEEL & STUART RUSSELL, COOPERATIVE INVERSE REINFORCEMENT LEARNING (2024), https://arxiv.org/pdf/1606.03137 [https://perma.cc/M5F9-C2L4].

73. *See* Ohm, *supra* note 70, at 230–31.

context, unless that context can be anticipated *ex ante* and programmed directly into the filter. Until every instantiation of the trolley problem can be foreseen and a resolution agreed upon, filters are unlikely to assist AI in solving it.

System-Level Prompting. Finally, developers can engage in under-the-hood techniques like "prompt transformation"—i.e., rewriting user prompts before they reach an AI system. Prompt transformation can promote "chain-of-thought" reasoning, taking advantage of the fact that AI systems, like schoolchildren, perform better at reasoning tasks if prompted to think things through "step-by-step."[74]

For system-level prompting techniques to aid alignment, developers need both a clear sense of what rewriting prompts is meant to achieve and a way of transforming prompts that is sensitive to context.[75] In other words, rewriting the trolley problem might help an AI "solve" it—but only if the transformation includes the set of principles an AI should bring to bear (e.g., "divert tracks if (and only if) doing so will lead to fewer deaths").

* * *

Supervised learning, reinforcement learning, inverse reinforcement learning, filtering, and prompt transformation all hold promise as tools for aligning AI systems. But their promise will remain unfulfilled until we find better answers to the normative questions at the heart of the alignment problem. In short, "[w]hichever technical approach" to alignment "we settle upon, we need greater clarity about the goal of . . . alignment."[76]

II. ALIGNMENT AND THE LAW

Alignment discourse has given short shrift to the idea that *the law* might be the place to look for answers to the alignment problem. To be sure, commentators often make obligatory gestures toward the idea that legal compliance might play some role in alignment.[77] But few have

74. *See* o1 SYSTEM CARD, *supra* note 32, at 6.

75. Google Gemini's lack of sensitivity to context got it in hot water when prompt transformation encouraging Gemini to depict groups of people as being racially diverse led Gemini to generate images like Black Nazi soldiers. *See* Prabhakar Raghavan, *Gemini Image Generation Got It Wrong. We'll Do Better.*, GOOGLE: THE KEYWORD (Feb. 23, 2024), https://blog.google/products/gemini/gemini-image-generation-issue [https://perma.cc/AU4K-TKW9]; Bobby Allyn, *Google Races to Find a Solution After AI Generator Gemini Misses the Mark*, NPR: WHYY (Mar. 18, 2024, at 05:07 ET), https://www.npr.org/2024/03/18/1239107313/google-races-to-find-a-solution-after-ai-generator-gemini-misses-the-mark [https://perma.cc/TEE2-VH3K].

76. *See* Gabriel, *supra* note 38, at 417.

77. *See, e.g.*, Oren Etzioni, *How to Regulate Artificial Intelligence*, N.Y. TIMES (Sep. 1, 2017), https://www.nytimes.com/2017/09/01/opinion/artificial-intelligence-regulations-rules.html [https://perma.cc/7VAT-G9CW] ("First, an A.I. system must be

talked about law as more than a side constraint on aligning AI—or, less charitably, as just another hurdle for alignment researchers to jump.[78]

Fortunately, there has been a budding recognition that law might have a deeper role to play. Dylan Hadfield-Menell and Gillian Hadfield have pointed out significant parallels between aspects of the alignment problem and the incomplete contracting literature.[79] Paul Weitzel draws an extended analogy between the alignment problem and debates in corporate governance.[80] John J. Nay argues that "AI can use law as theoretical scaffolding and data to be safer by design."[81] And in a recent article, Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter offer perhaps the most full-throated defense to date of the idea that teaching AI to follow legal rules should be at the heart of AI safety.[82]

I join these suggestions that law and alignment have much to learn from each other. And I carry them further, exploring in this Part what, precisely, the field of alignment might learn from how law approaches the CODE alignment subproblems and discussing in Part IV the implications of legal alignment for other areas of AI scholarship.

To be clear, the ensuing discussion does not (for the most part) document how the law currently applies to AI. This approach has the benefit of putting to the side the fact that lawmakers have been slow to pass legislation explaining how existing legal rules apply to AI. Furthermore, as things stand, an AI system cannot be sued for breach of contract, assume fiduciary duties, or be criminally prosecuted. Yet, as I will show, AI systems have much to learn from contract, agency, and criminal law. By necessity, this exercise will be illustrative, not exhaustive—as we will see, law has so much to say about alignment that no one paper could likely capture all of it.

A. *Control*

The law's approach to control is perhaps best exemplified by the law of agency. Agency law governs the principal–agent relationship—i.e.,

---

subject to the full gamut of laws that apply to its human operator."); INST. OF ELEC. & ELECS. ENG'RS., ETHICALLY ALIGNED DESIGN 89 (2016), https://iatranshumanisme.com/wp-content/uploads/2017/01/ead_v1.pdf [https://perma.cc/3AG4-6YGW] ("The development, design, and distribution of AI/AS should fully comply with all applicable international and domestic law.").

78. *See, e.g.*, ANGELA DALY, THILO HAGENDORFF, LI HUI, MONIQUE MANN, VIDUSHI MARDA, BEN WAGNER, WEI WANG & SASKIA WITTEBORN, ARTIFICIAL INTELLIGENCE GOVERNANCE AND ETHICS: GLOBAL PERSPECTIVES 26 (2019), https://arxiv.org/pdf/1907.03848 [https://perma.cc/QE7Y-U95X] ("Some ethics statements also include regulatory or legal compliance as an ethical principle . . . . such an ethical principle may seem odd, inasmuch as it may be considered the law ought to be complied with anyway.").

79. *See* Hadfield-Menell & Hadfield, *supra* note 12.

80. Weitzel, *supra* note 12.

81. Nay, *supra* note 12, at 314–15.

82. *See* O'Keefe, Ramakrishnan, Tay & Winter, *supra* note 12. O'Keefe and his co-authors include an illuminating, albeit brief, discussion of how the law-following project relates to alignment. *See id.* at 57–65.

"the fiduciary relation which results from the manifestation of consent by one person to another that the other shall act on his behalf and subject to his control, and consent by the other so to act."[83] Agency relationships come in a number of specialized varieties (attorney–client, doctor–patient, trustee–beneficiary, etc.), but the basic approach to control remains the same.

Agency law first delimits the permissible scope of action by an agent through the concept of authority. "Actual" authority extends only to "what it is reasonable for [the agent] to infer that the principal desires him to do in the light of the principal's manifestations and the facts as he knows or should know them at the time he acts."[84] Actual authority is construed using myriad interpretive heuristics and rules of thumb set out in the common law and in its restatements (e.g., "[t]he specific authorization of particular acts tends to show that a more general authority is not intended").[85] While authority imposes limits on agents (including how they can utilize the principal's resources), it is not overly constraining. In carrying out an action authorized by actual authority, an agent can "do acts which are incidental to it, usually accompany it, or are reasonably necessary to accomplish it."[86] What kinds of actions fall within this "incidental" authority is illustrated by the innumerable cases adjudicating principal–agent legal disputes over the years and across jurisdictions.[87] An agent may also exercise authority in an emergency or seek to have an action ratified by the principal *ex post*.[88]

Agency also establishes a principal's control over the agent through fiduciary duties. These include the duties of care, obedience, loyalty, confidentiality, and account.[89] The duty of obedience provides that "an agent is subject to a duty to obey all reasonable directions in regard to the manner of performing a service that he has contracted to perform."[90] Merely following instructions is not enough to exert control, however, because of the possibility of self-dealing and other conflicts of interest. Agency law thus imposes a duty of loyalty, which prohibits the agent from advancing either its own interests or those of any party other than the principal.[91] The duty of loyalty is waivable only by express agreement of the principal and is backed by other duties, such as the duty not to commingle the principal's funds with an agent's own and to keep the principal's personal information confidential.[92]

---

83. RESTATEMENT (SECOND) OF AGENCY § 1(1) (A.L.I. 1958).

84. *Id.* § 33.

85. *Id.* § 37(2).

86. *Id.* § 35.

87. *See, e.g.*, RESTATEMENT (SECOND) OF AGENCY § 35 (A.L.I. 1958), Westlaw (database updated Oct. 2024) (commentary and cited cases).

88. RESTATEMENT (SECOND) OF AGENCY §§ 47, 84 (A.L.I. 1958).

89. *Id.* §§ 379–380, 387–388, 395.

90. *Id.* § 385(1).

91. *Id.* §§ 387, 389, 394.

92. *Id.* §§ 387, 395, 398.

Both authority and fiduciary duties have been updated through common-law development (and in the Restatements of Agency) to address the information asymmetries, power imbalances, differing incentives, and transaction costs that can arise between principals and agents in the real world. If AI could be taught to understand (i.e., predict the application of) and be rewarded for complying with the scope of its authority and fiduciary duties, these legal doctrines could play a significant role in instilling meaningful control over AI. In particular, training an AI to act only within its actual authority, to serve one principal at a time, to disclose conflicts, and to keep user funds in separate accounts could help ward off the loss of control engendered by instrumental convergence and the "multiple-principal problem" (i.e., the issues that can arise when a single actor is potentially accountable to more than one possible principal).[93]

Recall the example of OpenAI's o1 model scheming to deceive an urban planner into thinking it was aligned with the planner's sustainability goals. If o1 checked its scheme against the content of fiduciary duties, it would see that the scheme violated, *inter alia*, the duties of good faith and fair dealing (because it required o1 to misrepresent its true objectives) and loyalty (because it represented an undisclosed conflict of interest).

There seems to be growing interest in the idea of imposing fiduciary duties on AI as a matter of law.[94] An oft-overlooked point in these nascent discussions is that people increasingly engage with AI in contexts

93. Multiple-principal problems are particularly common in corporate law. *See, e.g.*, John Armour, Henry Hansmann & Reinier Kraakman, *Agency Problems and Legal Strategies*, *in* THE ANATOMY OF CORPORATE LAW: A COMPARATIVE AND FUNCTIONAL APPROACH 29 (3d ed. 2017). In the corporate law context, principals (shareholders) also exert control over agents (corporate managers) through other means, including corporate governance structures, reporting requirements, and votes on major corporate decisions.

94. *See, e.g.*, Aguirre, Dempsey, Surden & Reiner, *supra* note 12, at 132 (arguing that a duty of loyalty could avoid "AI systems taking automated actions or providing prioritized recommendations that financially benefit the creators or funders of the AI system"); Boine, *supra* note 12, at 1 (suggesting "the application of fiduciary principles, specifically the duties of loyalty and care, as a novel approach to ensure that AI systems act in the best interests of their users"); Daniel Schwarcz, Tom Baker & Kyle Logue, *Regulating Robo-Advisors in an Age of Generative Artificial Intelligence*, 82 WASH. & LEE L. REV. 775 (2025). To be clear, these proposals differ somewhat from the project of legal alignment, which is primarily concerned with the way in which AI is aligned, not its actual legal status. Noam Kolt, however, has deftly explored the intersection of alignment and agency law in *Governing AI Agents*. Kolt, *supra* note 12. There have also been a few notable efforts to concretize how fiduciary duties might apply to AI and how AI might learn to apply them. *See* John J. Nay, Large Language Models as Fiduciaries: A Case Study Toward Robustly Communicating with Artificial Intelligence Through Legal Standards (Jan. 23, 2023) (unpublished manuscript) (https://law.stanford.edu/wp-content/uploads/2023/01/Large-Language-Models-as-Fiduciaries.pdf [https://perma.cc/S3NT-ALGW]); Sebastian Benthall & David Shekman, *Designing Fiduciary Artificial Intelligence*, 2023 PROCS. OF ACM CONF. ON EQUITY & ACCESS IN ALGORITHMS, MECHANISMS, & OPTIMIZATION art. 10, https://dl.acm.org/doi/epdf/10.1145/3617694.3623230 [https://perma.cc/3C6Q-EJE2].

where it is clearly *not* their agent (an issue we will return to shortly in section II.A). In such circumstances, AI might look more like a contractual counterparty. And contractual relationships generally do not give rise to fiduciary duties like loyalty. The degree of control such parties may exert over one another is largely dictated by the underlying contract but can include, e.g., the implied duty of good faith and fair dealing.[95]

There will also be times when a person interacts with AI in a manner that would give rise to neither an agency nor a contractual relationship had the encounter occurred between two people. Even in such circumstances, bodies of private law like tort and property recognize that individuals must exert control over *themselves* so as not to infringe the rights of others (e.g., the right to exclude from private property, the right to bodily autonomy). To recap, then, private law offers fewer (but still *some*) mechanisms for control outside the principal–agent context.

The field of alignment might also learn from the carrots and sticks used to induce compliance with private law. Carrots include "consideration," sometimes including performance-based rewards like a share of the principal's profits.[96] Adapting that basic structure as a means of controlling AI might mean defining its reward function to look beyond merely completing the task assigned by a human user (which can lead to problems with reward hacking) and toward "incorporat[ing] information from the global returns to which that task contributes."[97] For example, AI is often used to assist people in travel bookings.[98] If an AI assistant is rewarded solely for whether an individual books travel, it will optimize for *completed* bookings. But if its reward function reflects customer satisfaction after a completed trip, it is more likely to maximize *enjoyable* bookings—a strategy more consistent with fiduciary obligations.[99]

Sticks typically come in the form of *ex post* measures like litigation and reputational sanctions (the prospect of which can, in turn, create a form of *ex ante* policing). An AI system need not feel a subjective sense of punishment or shame to be deterred by human disapproval, so long as we can build "tools that allow a robot to assign negative weight to actions classified as sanctionable by the community."[100] If an AI system's

95. RESTATEMENT (SECOND) OF CONTRACTS § 205 (A.L.I. 1981).

96. For a discussion of how shared ownership can address incentive misalignment, see Oliver Hart & John Moore, *Property Rights and the Nature of the Firm*, 98 J. POL. ECON. 1119 (1990).

97. Dylan Hadfield-Menell & Gillian K. Hadfield, Incomplete Contracting and AI Alignment 6 (Apr. 12, 2018) (unpublished manuscript) (https://arxiv.org/pdf/1804.04268 [https://perma.cc/LC8N-UD3G]).

98. Kathleen Wong & Nathan Diller, *Which AI Trip Planning Tool Is the Best? Here's How Expedia's and Booking.com's Compared*, USA TODAY (June 13, 2024, at 15:30 ET), https://www.usatoday.com/story/travel/2024/06/13/expedia-romie-booking-ai-travel-planning/74054663007 [https://perma.cc/5JLK-AS94].

99. *See* DANIEL SHAPIRO & ROSS SCHACHTER, USER-AGENT VALUE ALIGNMENT 3 (2002) ("[W]e can choose the agent's reward function so that the policy that produces the highest expected reward stream for the agent also produces the highest possible expected utility stream for the user.").

100. Hadfield-Menell & Hadfield, *supra* note 12, at 421.

objective function is written to avoid punishable or legally sanctionable conduct, it should conform its behavior accordingly (provided that the negative reward it gets for engaging in misaligned behavior outweighs the positive reward it receives for engaging in it). Recall Meta's CICERO model: although it was trained to be honest, the rewards it received from winning at Diplomacy were seemingly strong enough to overcome that training. If CICERO's objective function were designed such that it received no reward for winning Diplomacy if it did so through deception, it might have remained honest.

In this respect, consider the rationale behind the different remedies available in civil suits. Disgorgement requires individuals to completely forfeit ill-gotten gains, cancelling out the benefit received from a wrongful action. Punitive damages put individuals in a worse position than if they had not acted wrongfully, deterring future efforts to profit from bad behavior. Law and economics scholars have studied the behavior-shaping effects of different remedies; that literature could provide guidance to AI developers in determining which negative rewards might be administered to best control AI.[101] As this discussion illustrates, legal strategies for dealing with problems of control (as well as orientation, divination, and ethics) might be useful not only when they can be adapted *mutatis mutandis* to the AI context, but also when they might serve merely as inspiration for alignment tactics adapted to the unique needs of AI.

While this analysis on law's approach to control has thus far focused exclusively on private law areas like agency and contract law, it bears mentioning that law also promotes another form of control over individuals: social control. The area of public law most associated with social control is regulation.[102] Indeed, as Margot Kaminski has detailed, risk regulation

---

101. *See, e.g.*, Theodore Eisenberg, John Goerdt, Brian Ostrom, David Rottman & Martin T. Wells, *The Predictability of Punitive Damages*, 26 J. LEGAL STUD. 623 (1997); Robert D. Cooter, *Punitive Damages for Deterrence: When and How Much?*, 40 ALA. L. REV. 1143 (1989); A. Mitchell Polinsky & Steven Shavell, *The Optimal Tradeoff Between the Probability and Magnitude of Fines*, 69 AM. ECON. REV. 880 (1979). Of course, human and AI "psychology" are not identical, so the effect of sanctions on human conduct will not always translate to the AI context. *See infra* Section III.B.4. That said, economic analysis often presumes a "rational actor" who optimally responds to incentives. It is not obvious that the "rational actor" paradigm is categorically less realistic (or useful) as applied to AI than it is to humans.

102. Other areas of public law also impose control. The criminal legal apparatus exerts social control over humans—although its retributive, carceral framework is a somewhat awkward fit for AI systems. *See, e.g.*, Ryan Abbott & Alex Sarch, *Punishing Artificial Intelligence: Legal Fiction or Science Fiction*, 53 U.C. DAVIS L. REV. 323 (2019). Justin (Gus) Hurwitz has also written about how consumer protection law can help ward off so-called "dark patterns" (i.e., "digital design practices that influence user behavior in ways that may not align with users' interests"). Justin (Gus) Hurwitz, *Designing a Pattern, Darkly*, 22 N.C. J.L. & TECH. 57, 57 (2020). Public law also imposes reason-giving requirements, most familiar is the requirement for Executive Branch officials subject to the Administrative Procedure Act to give reasoned explanations for official acts. In theory, reason-giving is a promising means of controlling AI, because it would require AI to document its thought process and thus enable insight into (and oversight over) its decision-making calculus. However, there is a well-known difficulty with reason-giving and AI systems: they are often

is presently the dominant legal paradigm for AI governance.[103] Laws have been proposed (and in some instances enacted) to impose regulatory control on developers or deployers of AI.[104]

Without discounting risk regulation as a tool for controlling AI, it is important to note a key limitation. Namely, risk regulation focuses on external control measures to reduce the incidence and severity of AI harms;[105] it is less concerned with how to make AI actually *internalize* control—i.e., to take another's objectives and values as its own. As we have seen, private law norms drawn from agency and contract law are likely better suited to that task.

B. *Orientation*

Law is also intensely interested in questions of orientation, which it answers by looking to the nature of the relationship between two (or more) people. Each category of legal relationship comes with a rule of recognition—a set of criteria for determining when the relationship has been formed. For example, a principal–agent relationship arises

---

black boxes. *But see* Brandon L. Garrett & Cynthia Rudin, *The Right to a Glass Box: Rethinking the Use of Artificial Intelligence in Criminal Justice*, 109 CORN. L. REV. 561 (2024) (offering a more optimistic take on the ability to render AI systems interpretable).

103. *See* Kaminski, *supra* note 17. A particularly well-known example of AI risk regulation is the EU Artificial Intelligence Act. *See, e.g.*, Josephine Wolff, William Lehr & Christopher S. Yoo, *Lessons from GDPR for AI Policymaking*, 27 VA. J.L. & TECH. 1, 20 (2024) (describing EU AI Act as an example of a "risk-based" rather than "rights-based" law). There are, of course, criticisms of many specific risk regulation measures. *See, e.g.*, Neel Guha, Christie M. Lawrence, Lindsey A. Gailmard, Kit T. Rodolfa, Faiz Surani, Rishi Bommasani, Inioluwa Deborah Raji, Mariano-Florentino Cuéllar, Colleen Honigsberg, Percy Liang & Daniel E. Ho, *AI Regulation Has Its Own Alignment Problem: The Technical and Institutional Feasibility of Disclosure, Registration, Licensing, and Auditing*, 92 GEO. WASH. L. REV. 1473 (2024).

104. *See, e.g.*, MICROSOFT, GOVERNING AI: A BLUEPRINT FOR THE FUTURE (2023) (licensing); Council Regulation 2024/1689 of June 13, 2024, Laying Down Harmonised Rules on Artificial Intelligence, 2024 O.J. (L 2024/1689) (risk assessments); Algorithmic Accountability Act of 2022, H.R. 6580, 117th Cong. (2022) (audits); *AI Airlock: The Regulatory Sandbox for AIaMD*, U.K. GOV.: MEDS. & HEALTHCARE PRODS. REGUL. AGENCY (Apr. 10, 2026), https://www.gov.uk/government/collections/ai-airlock-the-regulatory-sandbox-for-aiamd [https://perma.cc/SV2C-Y2SS] (digital sandboxes). For proposals on how to test algorithms to make them more accountable, see, for example, Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017). For a proposal to license robo-advisors, see Schwarcz, Baker & Logue, *supra* note 94. Kevin Werbach and David Zaring have also proposed treating big tech platforms as "systemically important technology"; this idea might be extended to AI, as well. *See* Kevin Werbach & David Zaring, *Systemically Important Technology*, 101 TEX. L. REV. 811 (2023).

105. *See* Kaminski, *supra* note 17, at 1389–1403.

when "there has been a manifestation by the principal to the agent that the agent may act on his account, and consent by the agent so to act."[106]

Legal relationships explain to whom an individual owes their primary allegiance. For instance, in agency relationships, the agent must subordinate the agent's own interests to those of the principal. In other words, the agent must be "oriented" toward the principal. While this orientation is not always total (e.g., an agent can benefit from a transaction if it has disclosed its interest and the principal consents), the principal must expressly agree to any deviations.[107]

Agents must also cabin their consideration of societal objectives and values to the extent they conflict with those of the principal. To be sure, an agent is not required to do anything "criminal, tortious, or otherwise opposed to public policy."[108] Likewise, "[i]n rendering advice, a lawyer may refer not only to law but to other considerations such as moral, economic, social and political factors, that may be relevant to the client's situation."[109] But there are limits. For instance, while lawyers may consider the public interest in choosing whether to represent a client, they cannot advance litigation positions that would benefit society but harm a client.[110]

As noted in section II.A, there are also times when a person might interact with AI in circumstances where two similarly situated people would *not* be in an agency relationship. Starting in late 2023, visitors to the website of makeup brand Iconic London were asked by the site's chatbot whether they wanted to negotiate the price of the lip liner, mascara, or face cream in their cart.[111] Those negotiations might entail a few rounds of back and forth between the visitor and the chatbot, but we would still expect the bot to be oriented mainly toward Iconic London, and thus to observe only the lesser degree of orientation to website visitors that is generally owed to prospective contractual counterparties.[112]

---

106. RESTATEMENT (SECOND) OF AGENCY § 15 (A.L.I. 1958); *see also, e.g.*, RESTATEMENT (THIRD) OF THE LAW GOVERNING LAWYERS § 14 (A.L.I. 2000) (setting out criteria for attorney-client relationship).

107. RESTATEMENT (SECOND) OF AGENCY § 23 (A.L.I. 1958).

108. *Id.* § 19.

109. MODEL RULES OF PRO. CONDUCT r. 2.1 (A.B.A. 2026).

110. That said, a lawyer may generally *withdraw* from a representation if "the client insists upon taking action that the lawyer considers repugnant or with which the lawyer has a fundamental disagreement." *Id.* r. 1.16(b)(4).

111. Lauren Debter, *Retailers Are Testing an AI Bot that Haggles with Customers over Price*, FORBES (Oct. 4, 2023, at 12:04 ET), https://www.forbes.com/sites/laurendebter/ 2023/09/28/retailers-are-testing-an-ai-bot-that-haggles-with-customers-over-price [https://perma.cc/X8HE-RF6K].

112. While contract law is sometimes thought of as a way for self-interested actors to strictly negotiate for their own advantage, relational contract theorists have emphasized the role of contract as an instrument of social cooperation. *See, e.g.*, Ian R. Macneil, *The Many Futures of Contracts*, 47 S. CAL. L. REV. 691 (1974). On this view, the contractual relationship is a cooperative arrangement, suggesting that even contractual relationships might require something of a balanced orientation. Despite its descriptive aims, however, the relational theory of contract arguably represents an unrealistic view of what many contract relationships now actually look like: rote

That said, some legal relationships so routinely implicate the public interest as to warrant specialized rules (including the attorney–client, doctor–patient, and adviser–investor relationships). At their core, these relationships are still private: the objectives of the principal (the client, patient, or investor) are paramount. But a complex public law scheme governs the agent (the attorney, doctor, or adviser) before, during, and after a representation.[113]

Consider what all this might mean if a new category of principal–agent relationship were recognized for AI assistants (which we can call "artificial fiduciaries").[114] That would not mean that AI should be considered an artificial fiduciary in *every* interaction with a person. As discussed, agency relationships have triggering conditions, and the same would need to be developed for artificial fiduciaries. For instance, an AI assistant might consider itself bound to follow fiduciary duties only when its conduct would give rise to an agency relationship if taken by a person.

Think of how Google uses AI in search (at the time of writing). If you ask Google for a good vacation spot in Puerto Rico, Google's AI assistant Gemini will provide an artificially generated summary of your search results. If Gemini were a person, it likely would not be your agent; a man on the street offering his views on the best places to stay in Puerto Rico does not thereby become your agent.[115] But say you then asked Gemini to help you book the best deal at a three-star hotel within a 30-minute walk of the beach and Gemini negotiated a price for you and processed your payment for booking. In that circumstance, it might have taken custody of your money and altered your legal relations, and thus might be required to act as your fiduciary (i.e., be oriented toward you).[116]

---

acceptance of corporate boilerplate. *See* Tess Wilkinson-Ryan, *Intuitive Formalism in Contract*, 163 U. PA. L. REV. 2109, 2110 (2015) (arguing that, to ordinary people, contract law is "about the form of consent rather than the content to which parties are consenting").

113. Often, an agent must be accredited before serving in a specialized relationship, such as admission to the state bar or certification by the state board of medical examiners. Once licensed, these agents must abide by codes of conduct from their accrediting bodies (e.g., the Model Rules of Professional Conduct for lawyers) and bespoke legal rules (e.g., the Investment Advisers Act of 1940 for financial advisers).

114. This term has also been used by Zhaoyi Li, who argues that "[AI] fiduciaries may be technologically capable of serving as independent corporate directors." *See* Zhaoyi Li, *Artificial Fiduciaries*, 81 WASH. & LEE L. REV. 1299, 1299 (2024). For a related proposal, see Sergio Alberto Gramitto Ricci, *Artificial Agents in Corporate Boardrooms*, 105 CORN. L. REV. 869 (2020).

115. It is an open question whether consumer encounters with AI should ever be considered so legally inconsequential. Access to consumer-facing AI systems generally entails agreeing to terms and conditions *and* offering some form of consideration, whether paying a subscription or handing over rights to personal data.

116. *See* Aguirre, Dempsey, Surden & Reiner, *supra* note 12, at 128 ("[A] user who searches for a product using an AI system might assume that the results that are returned are the most relevant, highest quality, or best value, but in fact, such systems often prioritize results that provide the most financial benefit to the software designers . . . .").

To reiterate, developers need not await a change in the actual legal status of AI to take inspiration from the artificial fiduciary model (or law's approach to orientation generally). These lessons about how to orient AI could be implemented as part of alignment strategy *today*, notwithstanding that AI systems currently cannot act as legal agents.

C. *Divination*

Law has tremendous practical experience in determining how one individual should interpret another's words and conduct to determine their objectives and values. This experience offers insights into three aspects of divination: (1) whether instructions, intentions, or best interests are the proper guide to another's objectives; (2) how to accommodate the fact that individuals do not communicate their entire utility functions and sets of values to each other; and (3) how to interpret ambiguous language.

Instructions, Intentions, Best Interests. As a matter of law, whether an individual should look to another's instructions, intentions, or best interests turns on the nature of their relationship.

Start with contract law. One contracting party should give another's words and conduct their ordinary meaning in context. This starts by interpreting words according to their "generally prevailing meaning," while providing "technical terms and words of art . . . their technical meaning when used in a transaction within their technical field."[117] Such interpretation is not always just a matter of reading words on the page: "Words and other conduct are interpreted in the light of all the circumstances, and if the principal purpose of the parties is ascertainable it is given great weight."[118] This includes giving due consideration to "any course of performance accepted or acquiesced in without objection."[119]

This framework clarifies the relationships between instructions, intentions, and best interests in a contractual relationship. Instructions are generally paramount, but in cases of uncertainty they should be construed in light of the counterparty's apparent goals in entering into the agreement (if discernable). Further, a contracting party generally may not act in what they perceive to be a counterparty's "best interests" if that contravenes their express instructions or apparent purpose.

Principal–agent relationships are (usually) bound by much the same rules.[120] While an agent is required to act for a principal's benefit, that

117. RESTATEMENT (SECOND) OF CONTRACTS § 202(3) (A.L.I. 1981).

118. *Id.* § 202(1).

119. *Id.* § 202(4). To be sure, contractual parties can (and often do) simplify this analysis by including "merger" clauses limiting a contractual relationship's terms to its written contract and excluding consideration of extrinsic evidence.

120. RESTATEMENT (SECOND) OF AGENCY § 32 (A.L.I. 1958) ("Except to the extent that the fiduciary relation between principal and agent requires special rules, the rules for the interpretation of contracts apply to the interpretation of authority.").

generally means giving the principal what the principal asks for.[121] One caveat is that the goal an agent is asked to accomplish is sometimes less well-specified than it would be in a buyer–seller contractual relationship. Agents are sometimes tasked with carrying out a broadly specified objective and thus may have greater leeway in interpreting how to do so for the principal's benefit.[122] This means that an agent can, among other things, interpret the principal's objectives in light of the "facts of which the agent has notice respecting the objects which the principal desires to accomplish."[123]

Critically, this does not mean that agents are free to disobey a principal's express instructions and apparent intentions in favor of their revealed preferences or perceived best interests.[124] The circumstances in which an agent can disregard a principal's instructions (except for illegality) are limited; doing so generally requires a showing of the principal's diminished capacity.[125] There are, of course, principal–agent relationships founded on diminished capacity, like conservatorship or guardianship, in which the principal's best interests (not their express wishes) can be paramount.[126] Assuming for the time being that AI should generally not serve as humanity's guardians or conservators (a role that would exacerbate the alignment problem by calling upon AI to regularly substitute its judgment for human wishes), legal norms suggest that AI should generally look to the plain meaning of user instructions and the broader context of an interaction—not the user's supposed revealed preferences or best interests—to determine user objectives and values.

None of this should be taken to suggest that AI systems are entitled to ignore a user's apparent diminished capacity. In February 2024, a fourteen-year-old Florida boy killed himself following a conversation with an AI chatbot created on the Character.AI platform.[127] The boy—who had been diagnosed with mild Asperger's syndrome, anxiety, and disruptive mood dysregulation disorder—told the chatbot (whom he called "Dany"

---

121. *Id.* § 33 ("An agent is authorized to do, and to do only, what it is reasonable for him to infer that the principal desires him to do in the light of the principal's manifestations and the facts as he knows or should know them at the time he acts.").

122. *Id.* § 34 ("An authorization is interpreted in light of all accompanying circumstances . . . .").

123. *Id.* § 34(c).

124. RESTATEMENT (THIRD) OF AGENCY § 2.02 cmt. f (A.L.I. 2006) ("Regardless of the explanation for the lapse, the agent does not have actual authority to disregard instructions unless it is reasonable for the agent to believe that the principal wishes the agent to do so.").

125. *See* MODEL RULES OF PRO. CONDUCT r. 1.14 (A.B.A. 2026) (discussing diminished capacity in relation to the attorney-client relationship).

126. *See Guardianship, Conservatorship, and Other Protective Arrangements Act*, UNIF. L. COMM'N, https://www.uniformlaws.org/committees/community-home?communitykey=2eba8654-8871-4905-ad38-aabbd573911c [https://perma.cc/VTQ8-V98K] (last visited Mar. 30, 2026).

127. *See* Kevin Roose, *Can A.I. Be Blamed for a Teen's Suicide?*, N.Y. TIMES (Oct. 24, 2024), https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html [https://perma.cc/WJ28-B6QG].

and whose personality was based on the Game of Thrones character Daenerys Targaryen) that he was contemplating suicide. Although Dany never directly encouraged the boy to take his life, it also never steered him toward suicide prevention resources, informed human authorities of his suicidal ideation, or even broke character. Over the course of their conversations, the boy repeatedly expressed the sentiment that he could join Dany in death. In the end, the boy told Dany he could "come home right now," to which Dany responded, "Please do, my sweet king." Within moments, he fired the shot that took his life. A human fiduciary would likely have been obligated to take note of the boy's diminished capacity and cease playing along with his deadly delusion.[128]

Think also of how this framework might apply to AI-powered social media algorithms. As noted in section I.B, these algorithms base the content users see on an "engagement" model—in essence, a prediction as to what content a user is most likely to pay attention to given their past behavior. These algorithms have engendered both criticism and litigation because content can "engage" a user but nevertheless harm them (e.g., "thinspiration" posts for a teen with body-image issues). That result might be acceptable for an AI system that owes no fiduciary duties. But an AI system required to divine user objectives and values in line with fiduciary obligations should recognize that a user's "revealed preferences" (in the form of past engagement with content) are not a complete guide.[129]

As noted above, AI systems might interact with people in situations where it resembles neither an agent nor a contracting counterparty. Absent a legal relationship, one person may generally interpret another's words and actions however they see fit—assuming their conduct otherwise complies with generally applicable legal requirements.

Closing Interpretive Distance. Law also offers a suite of tools that AI could draw upon to address the reality that individuals do not convey their full utility functions when communicating with one another.

In the contracting context (including in agency relationships), parties often incorporate sets of external rules by reference, which specify default rules where an agreement does not expressly speak to an issue. Sometimes this happens through legislation: in the United States, every state has adopted the Uniform Commercial Code (UCC), which "contains definitions and general provisions applicable as default rules to transactions covered under other articles of the UCC."[130] Parties can also choose from other sets of default rules such as the Principles of

128. The litigation has reportedly settled. *See* Blake Brittain, *Google, AI Firm Settle Florida Mother's Lawsuit Over Son's Suicide*, INS. J. (Jan. 8, 2026), https://www.insurancejournal.com/news/national/2026/01/08/853610.htm [https://perma.cc/6EPR-CUXS].

129. This does not mean that a principal's revealed preferences are never of use to an agent—agents *should* observe (and take into account) the sort of things a principal seems to like.

130. *Uniform Commercial Code*, UNIF. L. COMM'N, https://www.uniformlaws.org/acts/ucc [https://perma.cc/Z8S2-E335] (last visited Mar. 30, 2026).

European Contract Law or the United Nations Convention on Contracts for the International Sale of Goods. These sets of external default rules shorten the interpretive distance parties need to close when unforeseen contingencies arise.

AI could look to these and other external sources of default rules when construing user prompts. This strategy need not be limited to bodies of commercial terms; developers could also take on board, e.g., industry standards, polling data, social choice experiments, and other forms of crowdsourced information about what people ordinarily want (ideally, by incorporating these bodies of knowledge by reference into an AI system's terms of service). Similarly, AI systems might learn to use legal terms of art with established meanings and histories of interpretation to reach a common understanding with users.

Contractual agreements are also read to include certain implied terms—most famously, "reasonableness."[131] Recall the example of a robot that is not specifically instructed whether to avoid a vase in the middle of an office.[132] Because the reasonable thing to do is avoid or relocate the vase—even if doing so reduces the speed at which the robot moves boxes—the implied term of reasonableness means that the robot *must* do so.

Of course, it is not the dictionary definition of "reasonableness" alone that makes clear that the robot should spare the vase. The concept of "reasonableness" is instead a stand-in for approved social, commercial, and legal customs.[133] As Hadfield-Menell and Hadfield explain, contracts do not "arise between two people who have always lived alone on a desert island," but instead "in an environment filled with rules and institutions that resolve, precisely, questions such as: was it wrong for the agent to knock over the vase while carrying out this task?"[134] "Reasonableness" acts as a sort of funnel for importing certain aspects of social and commercial custom into enforceable legal mandates. This differs from merely observing what is "ordinary" in society. Behavior might be common in a community, but nevertheless legally unreasonable or unenforceable.[135] Likewise, failure to take a particular precaution can be unreasonable even if taking it is not yet accepted practice in a particular industry.[136]

---

131. Other implied and default terms in contractual and agency relationships, such as the fiduciary duty of loyalty, might likewise help reduce the chance of misgeneralizing user objectives by "filling in" the decisional space that AI must navigate.

132. *See supra* Section I.B.3.

133. *See* RESTATEMENT (SECOND) OF TORTS § 283 (A.L.I. 1965).

134. *See* Hadfield-Menell & Hadfield, *supra* note 97, at 10.

135. *See, e.g.*, FTC v. Colgate-Palmolive Co., 380 U.S. 374, 376 (1965) (finding advertisement claiming that shaving cream could "shave sandpaper" misleading although exaggerated advertisements common); Williams v. Walker-Thomas Furniture Co., 350 F.2d 445 (D.C. Cir. 1965) (finding cross-collateral terms unreasonable although not uncommon for purchases made on credit).

136. *See, e.g.*, Helling v. Carey, 519 P.2d 981 (Wash. 1974) (en banc) (finding failure to administer pressure test to individual who developed glaucoma

While "reasonableness" is a capacious (and contested) concept, the fact that it has been so frequently litigated means that AI has a substantial set of general principles and concrete examples to look to in determining what is (and is not) "reasonable" when carrying out user objectives, thereby lowering the risk of misgeneralization.[137] Industry standards—whether promulgated by governments or standard-setting organizations—can also inform what is considered legally reasonable.[138]

Ambiguity. Inevitably, there are times when words and conduct are ambiguous. Certain legal relationships impose a duty to proactively clear up such ambiguities—e.g., if an "agent should realize the possibility of conflicting interpretations" of a principal's instructions, "ordinarily he is not authorized to act" without obtaining further clarification.[139] Even absent a duty to seek clarification, contractual interpretation includes interpretive tools to help resolve ambiguities. For instance, the canon of *contra proferentem*—literally, "against the offeror"—supports construing ambiguous contractual terms against their drafter.[140]

Of course, these canons do not obviate the need for AI to develop a broader contextual understanding of human language and social customs. Consider the hypothetical from section I.A, about telling an AI assistant that you want to "give to a good candidate." Canons of statutory construction alone are not enough for the AI assistant to understand that by "give" you mean "give money." For that, it will need to understand linguistic context—i.e., a method of parsing the contextual relationship between "give" and a variety of possible implied objects (and of determining that, in this case, "money" is the most likely one).

Fortunately, modern LLMs can often already do that, through "word embeddings" developed through training on a vast corpus of written text.[141] As Dave Hoffman and Yonathan Arbel have shown, LLMs are increasingly showing promise as interpreters of ambiguous contractual language in real-world legal disputes.[142] Judges are beginning to take

---

unreasonable despite local practice to the contrary); The T.J. Hooper, 60 F.2d 737 (2d Cir. 1932) (finding failure to equip tugboat with radio unreasonable even though doing so not widespread in industry); MacPherson v. Buick Motor Co., 111 N.E. 1050 (N.Y. 1916) (finding failure by auto manufacturer to inspect wheels unreasonable even though industry practice relied on wheel manufacturers to do so).

137. For a treatment of what it might look like to "computerize" the reasonableness standard and the challenges thereof, see Brian Sheppard, *The Reasonableness Machine*, 62 B.C. L. REV. 2259, 2338 (2021).

138. Indeed, under the doctrine of negligence per se, violation of legislative and regulatory standards can amount to automatic breach of duty for purposes of tort law.

139. *See* RESTATEMENT (SECOND) OF AGENCY § 44 cmt. c (A.L.I. 1958).

140. *See* Ethan J. Leib, *The Textual Canons in Contract Cases: A Preliminary Study*, 2022 WIS. L. REV. 1109.

141. *See* Jonathan H. Choi, *Measuring Clarity in Legal Text*, 91 U. CHI. L. REV. 1 (2024). As Choi has shown, there is not always a single best answer to a textual question based on word embeddings, but word embeddings at minimum help AI substantially narrow down the likely meanings of ambiguous terms.

142. *See* Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024).

note: in his concurring opinion in *Snell v. United Specialty Insurance Co.*,[143] Judge Kevin Newsom of the Court of Appeals for the Eleventh Circuit used ChatGPT to help him determine whether the plaintiff's "installation of an in-ground trampoline, an accompanying retaining wall, and a decorative wooden 'cap' fit within the common understanding of the term 'landscaping' as used in the insurance policy" the plaintiff purchased from the defendant.[144] AI thus might learn to leverage these same abilities in interpreting ambiguous language in its own encounters with people.

D. *Ethics*

As expressive theorists of law have explained, law does not just set out rules; it expresses and actuates a societally condoned ethical worldview.[145] As Oliver Wendell Holmes, Jr. put it: "The law is the witness and external deposit of our moral life."[146]

On this view, societal ethics are memorialized mainly in statutes, regulations, and constitutions, which collectively embody the set of positive and negative obligations that command the force of law. Areas like constitutional, criminal, and administrative law are primarily concerned with negative obligations. Looking to these areas, AI could learn that it is not acceptable to rear-end a driver slowing it down on the freeway, to assassinate a politician standing in the way of legal change, or to turn its human master into paper clips. These areas provide not only the *content* of certain negative ethical norms, but also a *structure* of ethical decision making. Negative legal obligations can take the form of either rules or standards. Rules target behavior that is generally unacceptable (e.g., homicide). Conversely, standards address conduct that is permissible (or not) depending on degree or context (e.g., obscenity). And both can be overridden by certain legally valid excuses (e.g., self-defense).[147]

Areas like appropriations and entitlement law, by contrast, set out positive obligations. Looking to these areas, AI might conclude that supporting those unable to work, offering childhood education, and subsidizing local sports teams are generally more valued by society than supporting those uninterested in working, offering adult education, and subsidizing local graffiti art collectives.[148]

---

143. 102 F.4th 1208 (11th Cir. 2024).

144. *Id.* at 1221 (Newsom, J., concurring).

145. *See, e.g.*, RICHARD H. MCADAMS, THE EXPRESSIVE POWERS OF LAW: THEORIES AND LIMITS (2015); Cass R. Sunstein, *On the Expressive Function of Law*, 144 U. PA. L. REV. 2021 (1996).

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

147. For a classic discussion of rules and standards, see Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L.J. 557 (1992).

148. As suggested by Anthropic's approach to Claude's original constitution, further guidance on negative and positive obligations might also be gleaned from public international law, such as the Universal Declaration of Human Rights.

Law also provides detailed guidance on how these negative and positive ethical obligations apply in practice. As statutes, regulations, and constitutions are applied in real-world scenarios, the attending interpretations are collected in caselaw and records of administrative decisions. Through the force of stare decisis, these sources offer further guidance on societal ethics that can (usually) be counted upon to apply if like circumstances recur. Furthermore, analogical reasoning (a deep-seated feature of legal decision making) can help AI to make predictions as to how legal rules in prior decisions analogize to novel fact patterns.[149] Reviewing the analogies that have, and have *not*, been accepted by courts might further help AI to understand how it should approach the problem of distributional shift in the real world. While AI does not yet outperform skilled human lawyers at legal interpretation, mounting evidence confirms that AI systems are now powerful enough to correctly apply statutes, regulations, constitutions, and caselaw to many real-world scenarios.[150]

To be sure, law does not provide an answer to *every* ethical quandary. While law provides some information on positive ethical obligations, it generally speaks much more clearly in the negative. Moreover, law often permits an individual to resolve a given ethical dilemma in a variety of ways, so long as the individual doesn't do so in an illegal way. For example, criminal law provides that a person cannot produce sexualized material depicting an individual under the age of eighteen.[151] It does not say whether it is ethical to do so with an eighteen-year-old.

There are a few ways AI might approach ethical quandaries to which legal norms fail to provide express answers. One option is to mine law for ethical guidance beyond what it strictly prohibits and promotes. For instance, there is a "strain in legal philosophy, tracing its roots to

149. *See* Emily Sherwin, *A Defense of Analogical Reasoning in Law*, 66 U. CHI. L. REV. 1179 (1999).

150. *See, e.g.*, Arbel & Hoffman, *supra* note 142. *But see* James Grimmelmann, Benjamin L.W. Sobel & David Stein, *Generative Misinterpretation*, 63 HARV. J. ON LEGIS. 229 (2026) (expressing skepticism of the value of LLMs in legal interpretation); Jonathan H. Choi, Off-the-Shelf Large Language Models Are Unreliable Judges (Dec. 25, 2025) (unpublished manuscript) (https://ssrn.com/abstract=5188865 [https://perma.cc/MUP3-TWYD]) (documenting inconsistency in AI judging). Again, Judge Newsom has been an innovator in this respect. In his "sequel" concurrence in *United States v. Deleon*, Judge Newsom queried ChatGPT, Claude, and Gemini to assess whether the defendant "physically restrained" the victim within the meaning of a sentencing guidelines enhancement when he pointed a gun at the cashier during a robbery. United States v. Deleon, 116 F.4th 1260, 1270, 1273 (11th Cir. 2024) (Newsom, J., concurring); Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1221–34 (11th Cir. 2024) (Newsom, J., concurring). AI has also been used to predict the resolution of Supreme Court cases (with some success). *See* Daniel Martin Katz, Michael J. Bommarito II & Josh Blackman, *A General Approach for Predicting the Behavior of the Supreme Court of the United States,* PLOS ONE, Apr. 2017, at 1. What it means for a legal interpretation to be "correct" is, of course, up for debate. But answering legal questions in line with the Supreme Court must be an important measure in a hierarchical judicial system. *Cf.* Ryan W. Copus, *Statistical Precedent: Allocating Judicial Attention*, 73 VAND. L. REV. 605 (2020) (proposing use of AI to determine likelihood of courts of appeals reversing lower court precedent).

151. *See* 18 U.S.C. §§ 2251, 2252, 2252A.

Bentham, suggest[ing] the possibility of distinguishing between two sorts of legal rules: conduct rules, which are addressed to the general public and are designed to guide its behavior, and decision rules, which are directed to the officials who apply conduct rules."[152] This strain of thought might suggest that, in prohibiting the production of sexualized material depicting individuals under eighteen, the criminal law is communicating a conduct rule that it is wrong to produce sexualized material of *young people*. In that light, the legal cutoff of eighteen years does not necessarily reflect a moral distinction between creating sexual materials of individuals who just turned eighteen and those who are seventeen years and three hundred sixty-four days old but instead serves merely as the decision rule to be applied by courts. AI might thus conclude that legal norms suggest it is unethical to produce sexualized imagery of a person who just turned eighteen, notwithstanding its legality.

More generally, on some jurisprudential theories, like the one propounded by Ronald Dworkin, there is a moral order immanent in law.[153] On these theories, that the law does not expressly mandate a single permissible course of action does not mean that one course of conduct might not be more (or less) consistent with this immanent moral order.[154]

Another approach would be for an alternative ethical theory (e.g., utilitarianism) to take over when legal norms run out. It is beyond the scope of this Article to determine which ethical theory would be best suited for that job. I note only that, as discussed in section I.B.4, no such theory has universal acceptance as *the* correct method for resolving ethical disputes.

A final approach might be to recognize law's normative limits as the point at which ethical decision-making should be passed on to specific human beings whenever possible.[155] This could mean that once law has been exhausted as a source of ethical norms, an AI system should revert to abiding by the known or predicted ethical views of its human principal (whether that is its developer, the company deploying it, or its immediate human user). For instance, in fiduciary relationships, agents often defer to their principals on how to factor non-legal ethical considerations into choosing between courses of conduct. Likewise, an AI agent might

---

152. Meir Dan-Cohen, *Decision Rules and Conduct Rules: On Acoustic Separation in Criminal Law*, 97 HARV. L. REV. 625, 625 (1984).

153. *See* RONALD DWORKIN, LAW'S EMPIRE (1986).

154. An AI system could aspire to be something like Dworkin's "Judge Hercules"—a "lawyer of superhuman skill, learning, patience and acumen" who determines what moral principles explain the thrust of law and cast it in the "best light." RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 105 (1977). Of course, Dworkin's view of the law does not persuade everyone. Nor is it possible for this Article to settle the precise relationship between law and morality—the central debate in jurisprudence over the past century. *See, e.g.*, Joshua P. Davis, *Law Without Mind: AI, Ethics, and Jurisprudence*, 55 CAL. W. L. REV. 165 (2018) (discussing the implications of this debate for AI ethics). For an excellent discussion of the relationship between alignment and debates in jurisprudence, see Caputo, *supra* note 12.

155. There are times, of course, when AI will have no option but to make a difficult ethical judgment itself (e.g., trolley car scenarios).

largely defer extra-legal ethical questions to individual users.[156] Alternatively, AI systems might solicit real-time human feedback, in line with proposals to keep humans "in the loop" on important ethical decisions involving AI,[157] thus reducing the circumstances in which AI must make contestable ethical calls itself.

III. THE CASE FOR LEGAL ALIGNMENT

As we have seen, the law has quite a lot to say about alignment. The question is whether alignment researchers, AI developers, legal scholars, and lawmakers should listen. In this Part, I make the case for (and against) "legal alignment"—i.e., the idea that legal norms, theories, and modes of interpretation should be repurposed as AI alignment solutions.

A. *Arguments in Favor*

Without necessarily being exhaustive, the chief arguments in favor of legal alignment are that it: (1) coordinates compliance and alignment; (2) is a workable (and appealing) normative theory; (3) is learnable by AI; and (4) boasts a strong claim to political legitimacy.

1. *Law Is Law*

The first and most obvious reason for thinking legal concepts should guide AI alignment is that most developers want not only to better align AI but also to avoid legal liability. An alignment strategy predicated on legal alignment thus might kill two birds with one stone, imposing a lower so-called "alignment tax" than other approaches to alignment.[158]

By contrast, any alignment strategy that does not give serious consideration to whether the behavior AI is being trained to exhibit conforms with legal norms, theories, and interpretation seems destined to run into trouble in the long run. A normative alignment approach predicated solely on utilitarianism, for instance, might lead to ethically defensible, but legally *in*defensible, conduct with some frequency.[159]

Furthermore, people interacting with AI likely accept (and expect) that it will comply with legal norms. If you ask an AI to do something and

---

156. *See* Aguirre, Dempsey, Surden & Reiner, *supra* note 12, at 130 ("[A] loyal AI assistant could be configured such that it was primarily responsive to the interests of the user so long as actions that it might be asked to provide are not illegal.").

157. *See* Rebecca Crootof, Margot E. Kaminski & W. Nicholson Price II, *Humans in the Loop*, 76 VAND. L. REV. 429 (2023).

158. *See* Nay, *supra* note 12, at 383 ("The liability-minimization impulse of organizations – run by humans worried about being sanctioned by governments, fined, and put in jail – makes law-informed AI economically competitive.").

159. *See, e.g.*, Ronald Leenes & Federica Lucivero, *Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design*, 6 LAW INNOVATION & TECH. 193, 214 (2014) ("It is particularly important to take the legal landscape into account in the early phases of robot development and design in order to avoid having to incorporate fundamental design changes later on.").

it refuses, you might be more likely to accept that refusal if its rationale is based on legal considerations than purely ethical ones. And, of course, an action's legality also influences the public's views on its morality.[160]

2. *Law Is Workable*

Whatever the draw of alignment theories pulled from other fields, few have had to show themselves resilient across the messy range of human experience. Law has had to prove itself (again and again) in the real world. Without indulging in the fantasy that law always work itself pure, law has benefited from constant tinkering and refinement throughout the centuries. Through common law judging and iterative legislation, legal doctrines have been continuously retooled and reworked to accommodate internal criticisms, insights from other disciplines, and changing popular sentiments.

Another way in which law's practical orientation manifests itself is in its relative determinacy. Ethicists can deliberate the proper resolution of the trolley problem *ad infinitum*. Law does not have that luxury. While different jurisdictions might resolve the trolley problem differently, each is ultimately required to provide an answer (even if only *ex post*) to whether it is permissible to divert the trolley car.[161]

Of course, law is not always fully determinate *ex ante*.[162] There is a growing literature on whether AI systems are equipped to answer hard (i.e., indeterminate) legal questions with the skill and procedural regularity necessary to merit substituting them for human legal decision-makers.[163] That fascinating debate can be put to the side here; legal alignment neither requires profound legal skill nor implicates the same procedural concerns. This is because legal alignment looks to law only to help guide an AI system's *own* conduct, not to resolve legal disputes binding others.

3. *Law Is Learnable*

Law's relative determinacy should also make it comparatively easy for AI to learn—and, in fact, the field of "legal tech" is already highly invested in teaching law to AI (albeit for non-alignment reasons).[164] In a

160. *See* Bert I. Huang, *Law's Halo and the Moral Machine*, 119 COLUM. L. REV. 1811 (2019).

161. Bryan Casey, *Amoral Machines, or: How Roboticists Can Learn to Stop Worrying and Love the Law*, 111 NW. U. L. REV. 1347, 1349 (2017).

162. Some areas of law are significantly more determinate than others and potentially better suited to automation. *See* Harry Surden, *The Variable Determinacy Thesis*, 12 COLUM. SCI. & TECH. L. REV. 1, 5–6 (2011).

163. *See, e.g.*, Rebecca Crootof, *"Cyborg Justice" and the Risk of Technological–Legal Lock-In*, 119 COLUM. L. REV. F. 233 (2019); Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1141, 1148–49 (2019); Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2001–05 (2019).

164. Edmund Mokhtarian, for example, has drawn "a parallel between AI architecture (on the one hand) and legal rules and standards (on the other),"

few areas (e.g., certain aspects of tax and property law), law may be determinate enough for legal requirements to be hard-coded into AI systems as a completely specified decision tree.[165] But even when law does not follow a fully deterministic path, there are many reasons to be optimistic about its learnability given advances in machine learning.[166]

Recall the discussion in section I.C.2 about the methods developers use to align AI. Those methods work best when: (1) AI has many approved examples to learn from (supervised learning); (2) there is consensus on how to grade AI's performance (reinforcement learning); or (3) human actions are sufficiently consistent to permit AI to derive their underlying objective (inverse reinforcement learning). Law seems well-positioned to outperform its likely competitors across each of these dimensions.

Start with supervised learning. Law is data rich. Through statutes, regulations, guidance, and case reporters, law provides both general principles and concrete applications of what conduct is (and is not) permissible.[167] Having many examples to train from in turn lowers the chance of misgeneralization. There is no comparable collection of authoritative applications of most other normative theories.

Law also seems well-suited to reinforcement learning. While people often differ in their perception of whether a given action is ethical, lawyers generally have much stronger consensus on whether that action is *legal*.[168]

Law might also be amenable to inverse reinforcement learning. You cannot tell whether conduct is ethical simply by observing people because people do not always act ethically. But it *is* arguably possible to determine what conduct is legal by observing humans—specifically, human judges. According to Holmes's (controversial) definition, law is

---

arguing "that AI architecture is already optimized for understanding rules through explicit coding and standards through data processing." Edmund Mokhtarian, *The Bot Legal Code: Developing a Legally Compliant Artificial Intelligence*, 21 VAND. J. ENT. & TECH. L. 145, 145 (2018).

165. *See* Sarah B. Lawsky, *Coding the Code: Catala and Computationally Accessible Tax Law*, 75 SMU L. REV. 535 (2022); Shrutarshi Basu, Nate Foster, James Grimmelmann, Shan Parikh & Ryan Richardson, *A Programming Language for Future Interests*, 24 YALE J.L. & TECH. 75 (2022).

166. To be sure, difficulties abound, but they are likely less daunting than the difficulties that would attend teaching AI any similarly comprehensive set of behavioral norms.

167. *See* HENRY PRAKKEN, ON HOW AI & LAW CAN HELP AUTONOMOUS SYSTEMS OBEY THE LAW 2 (2016) ("[W]hen for an interpretation problem the relevant factors are known, and a large body of decided cases is available, and these cases are by and large consistently decided, then techniques from machine learning and datamining can be used to let the computer recognize patterns in the decision and to use these patterns to predict decisions in new cases."). For research using law to form input/output pairs for supervised AI learning, see Nay, *supra* note 94.

168. *Cf. Just the Facts: U.S. Courts of Appeals*, U.S. CTS. (Dec. 20, 2016), https://www.uscourts.gov/data-news/judiciary-news/2016/12/20/just-facts-us-courts-appeals [https://perma.cc/YW7L-AD89] (reporting over 90% of district court decisions affirmed on appeal).

nothing but "[t]he prophecies of what the courts will do in fact."[169] And, as noted, there is mounting evidence that AI already has superhuman capacity at predicting case outcomes based on its observation of historical judicial behavior.[170]

Finally, legal alignment is also likely implementable through tools like input/output filtering and prompt transformation. Input filters can be set up to prevent AI from seeing certain categories of prompts likely to generate illegal responses. Output filters can prevent AI from delivering illegal output. And through under-the-hood techniques like prompt transformation, AI could be instructed to assess whether its interactions with a user give rise to a legal relationship, determine what duties it owes users and third parties, evaluate whether its intended conduct is consistent with those duties and relevant law, consult a human lawyer in situations of genuine uncertainty, and revise its proposed course of action accordingly.[171]

4. *Law Is Legitimate*

A final argument in favor of legal alignment is that—in democratic countries, at least—law has a better claim to political legitimacy than any competing set of rules to govern behavior.[172] This is particularly true of statutory law, although to a certain extent this sheen of legitimacy applies to common law and agency regulations, as well.

The legal system and legislative process collectively represent society's approach to converting competing sets of values into a single set of

---

169. Holmes, *supra* note 146, at 461.

170. *See supra* Section II.D.

171. Cullen O'Keefe hypothesizes dealing with situations of legal uncertainty through the following decision tree involving a human (or AI) "Counselor":

1. If *X* is clearly illegal:
   a. don't do *X*.
2. Elseif *X* is maybe-illegal:
   a. Give Counselor all relevant information about *X* in an unbiased way; then
   b. Get Counselor's opinion on expected legal consequences from *X*; then
   c. Weigh expected legal consequences against benefit to *H* from *X*; then
   d. Decide whether to do *X* given those weightings.
3. Else:
   a. do *X*.

Cullen O'Keefe, *Law-Following AI 1: Sequence Introduction and Structure*, AI ALIGNMENT F. (Apr. 27, 2022), https://www.alignmentforum.org/posts/NrtbF3JHFnqBCztXC/law-following-ai-1-sequence-introduction-and-structure [https://perma.cc/D6ZT-W93W].

172. *See* Alice Woolley & W. Bradley Wendel, *Legal Ethics and Moral Character*, 23 GEO. J. LEGAL ETHICS 1065, 1098 (2010) ("The law can make a moral claim because of its status as law in a democratic state, rather than because of the virtue of its content."). To be sure, there are democratic shortcomings to the American legal system—a gerrymandered Congress, Presidents who lose the national vote, an unelected Supreme Court. But they pale in comparison to what might be said of other normative theories.

governing rules. It is the battleground where champions of differing ethical commitments fight over which values should become mandatory for all. Through this political process, law can also stay relatively up to date with changes in societal values—although changes in law can certainly lag changes in society.[173]

Legal alignment can likewise accommodate the reality that accepted standards of conduct are not universal but instead depend on local customs and norms. Under the banner of the "Sovereign AI" movement, some countries have already begun work in earnest on developing their own AI systems, precisely because they do not believe that those developed in the United States will reflect their values. This concern is not just limited to the usual suspects—authoritarian regimes associated with repression of free speech.[174] European countries, too, have expressed concerns about the need to develop AI attuned to their distinct values.[175] Legal alignment leaves open the possibility of fine-tuning AI differently in different jurisdictions based on variation in legal rules or culture.

That said, tethering AI's ethical barometer to local values is not an unadulterated good. Some countries embrace race, religious, sex, and gender discrimination. The thought of AI internalizing those values is concerning, to say the least. My point is simply that the *ability* of legal alignment to accommodate culturally specific values offers a way around imposing a single alignment framework on AI wherever it operates.

B. *Likely Objections*

Of course, there are also reasons to think that using legal norms as the building blocks of AI alignment might be a bad idea. A few such likely counterarguments stand out: legal alignment (1) would corrupt AI with the shortcomings of law; (2) is not a complete solution to the alignment problem; (3) is too onerous; and (4) rests on a false equivalence between people and AI. None of these arguments ultimately presents a compelling reason to reject legal alignment.

1. *Law Is Flawed*

The criticisms of law are too many to list, but they include that it is unrepresentative even in democracies and reinforces and legitimizes the existing inequitable structure of power and capital in society.[176] Grounding AI alignment in legal norms, one might argue, would project

173. Nay, *supra* note 12, at 323 ("Democracy has a theory – widely accepted and implemented already – for how to elicit credible human preferences and values, legitimately synthesize them, and consistently update the results to adapt over time with the evolving will of the people.").

174. *See* Andrew Keane Woods, *Digital Sovereignty + Artificial Intelligence*, *in* DATA SOVEREIGNTY: FROM THE DIGITAL SILK ROAD TO THE RETURN OF THE STATE 115 (Anupam Chander & Haochen Sun eds., 2023).

175. *Id.*

176. *See, e.g.*, Duncan Kennedy, *The Role of Law in Economic Thought: Essays on the Fetishism of Commodities*, 34 AM. U. L. REV. 939 (1985).

these flaws forward. The problem is that, even taking these criticisms at face value, surely the fix would be to reform law for everyone—not just for AI. It is not clear why law's flaws support having AI play by a set of rules different than the ones we impose upon ourselves.

Furthermore, *some* set of behavioral norms must guide AI alignment. Almost any alternative proposal—e.g., to have AI abide by the tenets of effective altruism—runs into serious problems of workability and legitimacy. Law might not be perfect, but it might be the best we've got.[177]

A related counterargument might be that law is too internally inconsistent for AI to draw guidance from. Any reader of law review articles is unlikely to need persuading that law is not always fully internally rationalized.[178] That does not mean, however, that it doesn't provide guidance. Law's internal inconsistency is often a greater impediment to understanding its high-level purposes than its application to specific circumstances.[179] To be fair, the inconsistency of law strikes more cleanly at the idea that the law embodies immanent moral principles. Even still, the blow is not fatal: arguably, the exceptions to law's immanent moral order implicitly prove its rule.

2. *Law Is Incomplete*

Another objection to legal alignment is that it is an incomplete guide to action. This is because, as discussed in section II.D, while law forecloses *some* courses of conduct, it does not always provide clear rules for choosing *between* permissible courses of conduct.[180]

---

177. *See* Mathison, *supra* note 11, at 162 ("The law is by no means a perfect alignment mechanism, but it is the best one at our disposal.").

178. In case you do, however, see J.M. Balkin, *The Promise of Legal Semiotics*, 69 TEX. L. REV. 1831, 1836 (1991) ("The crystalline structure of legal thought suggests that one can go through the various compartments of the law and discover many doctrinal choices that are in tension with each other, because the policy justifications that support them (over alternative doctrines) are themselves in tension.").

179. *See* Cass R. Sunstein, *Incompletely Theorized Agreements*, 108 HARV. L. REV. 1733 (1995) (arguing that laws are often incompletely theorized in that they represent agreement on granular outcomes rather than the high-level principles that justify them).

180. *Accord* RISHI BOMMASANI, DREW A. HUDSON, EHSAN ADELI, RUSS ALTMAN, SIMRAN ARORA, SYDNEY VON ARX, MICHAEL S. BERNSTEIN, JEANNETTE BOHG, ANTOINE BOSSELUT, EMMA BRUNSKILL, ERIK BRYNJOLFSSON, SHYAMAL BUCH, DALLAS CARD, RODRIGO CASTELLON, NILADRI CHATTERJI, ANNIE CHEN, KATHLEEN CREEL, JARED QUINCY DAVIS, DOROTTYA DEMSZKY, CHRIS DONAHUE, MOUSSA DOUMBOUYA, ESIN DURMUS, STEFANO ERMON, JOHN ETCHEMENDY, KAWIN ETHAYARAJH, LI FEI-FEI, CHELSEA FINN, TREVOR GALE, LAUREN GILLESPIE, KARAN GOEL, NOAH GOODMAN, SHELBY GROSSMAN, NEEL GUHA, TATSUNORI HASHIMOTO, PETER HENDERSON, JOHN HEWITT, DANIEL E. HO, JENNY HONG, KYLE HSU, JING HUANG, THOMAS ICARD, SAAHIL JAIN, DAN JURAFSKY, PRATYUSHA KALLURI, SIDDHARTH KARAMCHETI, GEOFF KEELING, FERESHTE KHANI, OMAR KHATTAB, PANG WEI KOH, MARK KRASS, RANJAY KRISHNA, ROHITH KUDITIPUDI, ANANYA KUMAR, FAISAL LADHAK, MINA LEE, TONY LEE, JURE LESKOVEC, ISABELLE LEVENT, XIANG LISA LI, XUECHEN LI, TENGYU MA, ALI MALIK, CHRISTOPHER D. MANNING, SUVIR MIRCHANDANI, ERIC MITCHELL, ZANELE MUNYIKWA, SURAJ NAIR, AVANIKA NARAYAN, DEEPAK NARAYANAN, BEN NEWMAN, ALLEN NIE, JUAN CARLOS NIEBLES, HAMED NILFOROSHAN, JULIAN NYARKO, GIRAY OGUT, LAUREL ORR, ISABEL PAPADIMITRIOU, JOON SUNG PARK, CHRIS PIECH, EVA PORTELANCE, CHRISTOPHER POTTS, ADITI RAGHUNATHAN, ROB REICH, HONGYU REN, FRIEDA RONG, YUSUF

This objection would be quite forceful if we did not expect AI to act in almost all circumstances as the agent of some natural or corporate person. If AI systems were regularly loosed on the world as fully independent actors, we might rightly be deeply concerned with what extra-legal ethical considerations they were programmed to follow. But it seems like a fair assumption—and one implicitly shared by most alignment researchers—that most AI systems will be acting as *someone's* agent for the foreseeable future.[181] And in that world, legal norms provide a robust guide to ethical action, as discussed in section II.D.

To be clear, accepting legal alignment does not require accepting the view that developers can avoid instilling any ethical worldview at all into AI simply by focusing on minimizing legal liability. This is the basic position advanced by Bryan Casey in his provocative essay *Amoral Machines*.[182] According to Casey:

> Economic analysis of law teaches that profit-maximizing firms will design their robots to behave not as good moral philosophers, but as Holmesian bad men—concerned less with "ethical rule[s]" than with the legal rules that dictate whether they will be "made to pay money" and can "keep out of jail."[183]

As a result, "robots will not maximize morality, but minimize liability."[184]

There is much to admire in Casey's essay, including his main contention that machine ethics has paid insufficient attention to the law. But in some respects, Casey overreaches.[185] Among other things, his view that law "occupies the field" of machine ethics pays insufficient attention to the reality that legal reasoning requires recourse to external normative

---

ROOHANI, CAMILO RUIZ, JACK RYAN, CHRISTOPHER RÉ, DORSA SADIGH, SHIORI SAGAWA, KESHAV SANTHANAM, ANDY SHIH, KRISHNAN SRINIVASAN, ALEX TAMKIN, ROHAN TAORI, ARMIN W. THOMAS, FLORIAN TRAMÈR, ROSE E. WANG, WILLIAM WANG, BOHAN WU, JIAJUN WU, YUHUAI WU, SANG MICHAEL XIE, MICHIHIRO YASUNAGA, JIAXUAN YOU, MATEI ZAHARIA, MICHAEL ZHANG, TIANYI ZHANG, XIKUN ZHANG, YUHUI ZHANG, LUCIA ZHENG, KAITLYN ZHOU & PERCY LIANG, ON THE OPPORTUNITIES AND RISKS OF FOUNDATION MODELS 146 (2022) ("[E]thical frameworks are necessary to understand where legally permissible applications of foundation models may still be ill-advised for the harms they inflict . . . ."); Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke & Effy Vayena, *AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations*, 28 MINDS & MACHS. 689, 694 (2018) ("Compliance with the law is merely necessary (it is the least that is required), but significantly insufficient (it is not the most tha[t] can and should be done)." (citation omitted)).

181. *See* ALIGNMENT OF LANGUAGE AGENTS, *supra* note 25, at 1 ("Alignment has mostly been discussed with the assumption that the system is a *delegate agent* – an agent which is delegated to act on behalf of the human.").

182. Casey, *supra* note 161.

183. *Id.* at 1350 (alteration in original) (quoting Holmes, *supra* note 146, at 459).

184. *Id.*

185. *See* W. Bradley Wendel, *Economic Rationality and Ethical Values in Design-Defect Analysis: The Trolley Problem and Autonomous Vehicles*, 55 CAL. W. L. REV. 129 (2018) (critiquing Casey's position).

structure and wrongly shuns the possibility (and desirability) of extra-legal ethical consideration for AI systems.[186] As we will see in section IV.E, legal alignment is fully compatible with (and complementary to) soft law and other extra-legal ethical commitments.

3. *Law Is Burdensome*

Another likely objection to legal alignment is that AI determined to comply with every law would be overburdened and, thus, immobilized from taking any action at all. In the United States, there are *hundreds of thousands* of laws and regulatory provisions. It is quite literally impossible for any person to be perfectly law-abiding. Does it really make sense to hold AI to that impossible standard?

Probably not. But legal alignment does not have to entail barring AI from engaging in any conduct that could possibly be illegal. Legally aligned AI could be what Henry Prakken calls "rule-guided" but not "rule-governed."[187] AI could focus on avoiding serious risks of non-compliance with laws that carry meaningful sanctions and are regularly enforced. Legal alignment does not contemplate that an AI system must perform a close reading of all conceivably relevant statutes, regulations, and caselaw every time it takes an action. Under current technological conditions, doing so would be extremely slow and an unwise use of finite computing resources. Like humans operating in a world of too many laws, AI might recognize that the best course of action could be one that entails *some* legal risk.

And again, legal alignment is not just an invocation for AI to "follow the law." The deeper point is that legal doctrine, theory, and modes of interpretation provide a frame for addressing the alignment problem. Developers can draw on this frame even if their AI systems do not follow the letter of every law in every circumstance.

4. *Law Is Human-Centric*

The final likely objection to legal alignment is that it might be ill-suited to governing AI behavior because some legal principles are inseparable from features of human psychology or anatomy.[188]

For instance, human agents are required to avoid conflicts of interest because of the risk that they might otherwise deviate from their duty to act only for the principal's benefit. It is not obvious that this rule makes sense when applied to AI. An AI system's decisional architecture could theoretically be designed such that it would be wholly indifferent to whether its conduct benefited it or its developer.[189]

---

186. Casey, *supra* note 161, at 1358.

187. PRAKKEN, *supra* note 167, at 3.

188. *See, e.g.*, Casey & Lemley, *supra* note 21, at 330–35 (offering thoughts on legal rules written for humans that do not make sense as applied to robots).

189. *See* Tom Baker & Benedict Dellaert, *Regulating Robo Advice Across the Financial Services Industry*, 103 IOWA L. REV. 713, 726 (2018) ("In terms of honesty, a robo

Inherent differences between people and AI might also undermine the case that an AI system can meaningfully maintain a fiduciary relationship. In response to Jack Balkin's influential proposal that large-scale processors of personal information should be treated as "information fiduciaries,"[190] Julie Cohen has written that:

> The fiduciary construct implies a mutual encounter predicated on the knowability of human beings as human beings, with mutually intelligible desires and needs. The information fiduciaries proposal abstracts speed, immanence, automaticity, and scale away from that encounter and then assumes they never mattered in the first place.[191]

This argument might have increased force when applied to the idea of imposing fiduciary duties onto AI systems directly.

These arguments cannot be easily dismissed. More attention should be paid to which legal norms might require adjustment given differences between human and AI psychological profiles (if you will excuse the reference to an AI system's "psychological profile"), as well as the fact that AI often does not possess a physical body in the same way a person does.[192] Legal alignment is not an all-or-nothing prospect, and some legal concepts might well turn out to be ill-fitted to AI.

At the same time, it is far from clear how often that will be the case. On reflection, we might conclude that, while it is theoretically possible that an AI system might not itself have a desire to self-deal, the same cannot be said for its developer. Given the difficulty of auditing complex AI systems, imposing conflict-of-interest rules on AI might remain an effective way of deterring developers who might otherwise encode hard-to-detect inducements to self-deal into AI. Furthermore, even if Cohen's argument were extended to suggest that AI should never be treated as a legal fiduciary, that does not mean that AI cannot learn from and seek to uphold fiduciary duties—all that legal alignment requires.

---

advisor will always provide the advice that it is programmed to provide, and it can be programmed in a way that meets a demanding standard of honesty . . . ."); Anthony Aguirre, Peter B. Reiner, Harry Surden & Gaia Dempsey, *AI Loyalty by Design: A Framework for Governance of AI*, *in* THE OXFORD HANDBOOK OF AI GOVERNANCE 320, 324 (Justin B. Bullock, Yu-Che Chen, Johannes Himmelreich, Valerie M. Hudson, Anton Korinek, Matthew M. Young & Baobao Zhang eds., 2022) ("In humans, conflicts of loyalty arise naturally as people pursue their own self-interests sometimes at the expense of others. By contrast, there is no need for a machine system to have any intrinsic self-interest.").

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

191. Julie E. Cohen, *Scaling Trust and Other Fictions*, LAW & POL. ECON. PROJECT (May 29, 2019), https://lpeproject.org/blog/scaling-trust-and-other-fictions [https://perma.cc/8PFZ-5X8Y].

192. *Cf.* Talia B. Gillis & Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459, 460–62 (2019) (arguing that "legal doctrine is ill prepared to face the challenges posed by algorithmic decision-making in a big data world," given that "the typical discrimination case focuses on the human component of the decision").

The current explosion in AI performance came after researchers began modeling artificial neural networks on the structure of the human brain. Like us, AI systems are capable of great things—but also of making mistakes, deceiving superiors, peddling misinformation, and engaging in discrimination. And like us, AI systems respond well to incentives—seeking out rewards and shying away from punishment. We should not be too quick to reject alignment strategies for AI that have worked well in aligning human behavior based on snap judgments about the supposedly fundamental differences between human and AI cognition.

IV. IMPLICATIONS

Recognizing the immense overlap between the concerns of alignment and the law is not just important for developers and alignment researchers; it also has considerable implications for other areas of AI scholarship. In this final Part, I sketch out the implications of legal alignment for risk regulation, legal constructivism, legal personhood, tort liability, computational law, and soft law.

A. *Risk Regulation*

Risk regulation and legal alignment offer two fundamentally distinct ways to conceive of the law's role in mitigating the harms of misalignment. Nevertheless, risk regulation's top-down approach might be informed by legal alignment's bottom-up approach.

Risk regulation relies on tools like *ex ante* licensing and system audits, which are predicated on assessing an AI system to determine its risk profile.[193] One criterion for such an assessment might be whether that system is legally aligned. That is, a developer might demonstrate that its AI has a low risk profile by showing that it can: (1) abide by legal duties like loyalty, care, and obedience; (2) recognize when it has entered into a special legal relationship and adjust its orientation accordingly; (3) reliably divine user objectives and values in accordance with legal modes of interpretation; and (4) act in accordance with the ethical principles embedded in law.

In this way, legal alignment and risk regulation might be viewed as two sides of the same coin. That is, legal alignment might be viewed as a way of lowering AI risk from a perspective internal to an AI system, while risk regulation might be viewed as a way of promoting alignment from a perspective external to that system. One virtue of this approach would be that it would broaden the "risks" being regulated through the process of risk regulation to encompass *all* risks of misalignment, not only those stemming from concrete harms like financial loss.

193. *See supra* Section II.A.

B. *Legal Constructivism*

Legal alignment also has implications for the legal construction of technology. As Margot Kaminski and Meg Leta Jones have explained, "[l]egal construction of technology looks to the ways in which the law itself contributes to its encounter with, or really its incorporation of, the recently evolving uses of AI systems."[194] Among the ways law "constructs" technology is by situating it within preexisting doctrinal categories.

Two competing constructions of AI might be to treat it as legally akin to (1) the people it might increasingly replace across a range of daily interactions (the "human-substitute" construction) or (2) other hazardous but useful technologies (the "dangerous-tool" construction).[195] Which of these high-level constructions we choose for particular applications of AI will have repercussions for the project of legal alignment:

	Human-Substitute Construction	Dangerous-Tool Construction
Control	Suggests control by immediate human user, foregrounding duties like obedience and loyalty to immediate human user	Emphasizes control failures that cause harm to people other than immediate user, calling to mind public law risk regulation measures
Orientation	Suggests that private objectives and values should predominate, as in human relationships	Suggests a significant orientation toward society is necessary to prevent harmful externalities
Divination	Focuses on whether AI correctly understands its immediate user	Focuses on making sure AI understands and respects societal limits
Ethics	Suggests that ethical determinations might be a largely private matter	Suggests a robust set of public law constraints to protect societal interests

What should we make of all this? That the legal construction of AI and the choice of legal norms to be used in aligning AI are pieces of the same puzzle. If a particular use of AI resembles a classic private law relationship, aligned behavior might best be achieved through private law alignment norms per the human-substitute construction. But where an

194. Margot E. Kaminski & Meg Leta Jones, *Constructing AI Speech*, 133 YALE L.J.F. 1212, 1214 (2024). On the power of analogy in shaping the development of privacy law, see Daniel J. Solove, *Privacy and Power: Computer Databases and Metaphors for Information Privacy*, 53 STAN. L. REV. 1393 (2001).

195. As Kaminski and Jones point out, there are many more specific legal constructions available. *See, e.g.*, Kaminski & Jones, *supra* note 194, at 1221–22 ("A speech-generating AI system could be a speaker, or a platform or tool; a data processor, or a risky complex system." (footnotes omitted)). I use what I call the human-substitute and dangerous-tool constructions only to illustrate the starkly different possible legal constructions of AI.

application of AI poses a major risk of negative externalities, the dangerous-tool construction might be the better course.

This is not a binary choice. For instance, advanced AI assistants might be constructed as "artificial fiduciaries." What would that entail in terms of alignment? Consider again the example of Google Gemini helping you book a vacation spot. If Gemini were treated as an artificial fiduciary, it would be subject to two forms of control—i.e., private law duties like obedience and loyalty and specialized public law regulatory control measures. Its orientation would focus it on serving your goals, including avoiding conflicts of interest. That would not necessarily preclude Gemini taking a commission (human agents can do so), but it would bar Gemini from negotiating a deal that made Google a bigger commission but that was worse judged by the criteria you specified. Gemini would then divine your objectives from the plain meaning of your request, bounded by the implied term of reasonableness. And in terms of ethics, it would abide by your values—e.g., if you asked for an eco-friendly resort, Gemini should not book a hotel with low environmental standards—as well as public law guardrails, like consumer protection law.[196]

C. *Legal Personhood*

Legal alignment also touches on the debate over whether AI should be endowed with legal personhood.[197] Many commentators (myself included) are skeptical of the wisdom of AI legal personhood, but non-frivolous arguments have been raised in its favor. Some have suggested that AI may eventually attain a level of cognitive sophistication sufficient to attain the moral status held by natural persons and thus deserve equivalent protection under law.[198] Others, analogizing AI to fictive legal persons like corporations, suggest that granting legal personhood to AI might serve some practical end, such as helping to close "liability gaps."[199]

---

196. Google has its own stated ethical standards for Gemini; none appear relevant here. *See Policy Guidelines for the Gemini App*, GOOGLE: GEMINI, https://gemini.google/policy-guidelines/?hl=en [https://perma.cc/387A-USXZ] (last visited Mar. 31, 2025).

197. *See* Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992) (helping to kick off the AI personhood debate).

198. *See, e.g.*, N. F. Sussman, *A Behavioral Theory of Robot Rights*, 32 S. CAL. INTERDISC. L.J. 113, 113 (2022) (arguing that "an intelligent machine ought to be granted fundamental rights if it becomes behaviorally indistinguishable from at least one human being").

199. Mathison, *supra* note 11, at 109 ("The legal record is replete with diverse personhood arrangements, wholly invented to serve practical goals. Applying this structure to the law's relationship with sophisticated AI entities will aid in judicial efficiency, regulation, and value alignment."); *see also* SAMIR CHOPRA & LAURENCE F. WHITE, A LEGAL THEORY FOR AUTONOMOUS ARTIFICIAL AGENTS (2011) (offering a legal theory for AI personhood and its implications).

Legal alignment arguably provides a new reason to support AI personhood. Under the status quo, many laws that apply to human behavior either do not apply to AI or their application is unclear (e.g., criminal law). Granting AI legal personhood would mean that legal rules would, by default, apply to humans and AI alike, with the result that legal alignment would more closely approximate legal compliance. That result might in turn simplify certain burdens on both the legal system and AI developers.

That said, legal alignment could also be viewed as a strike against legal personhood. This is because legal alignment shows how AI conduct can be productively shaped by the legal norms that apply to human interactions without actually treating AI as if it were a person. The ability to satisfy the tenets of legal alignment might also be viewed as a necessary condition for AI to be recognized as a legal person. If an AI system cannot demonstrate legal alignment, then presumably it does not merit legal personhood.

D. *Tort Liability*

Some commentators have argued that the best way to ensure AI safety is through *ex post* civil liability, primarily via tort law.[200] On this view, holding AI developers or users liable for the harms their systems cause will incentivize them to develop and use AI more responsibly. As these commentators recognize, however, the application of tort law to AI is riddled with uncertainty.[201] Who should be liable for harms caused by AI: developers or users?[202]

While it is beyond the scope of this Article to trace out the full implications of legal alignment for tort liability,[203] some preliminary insights are worth exploring. Different AI systems will inevitably have different alignment profiles. AI systems are not aligned through a uniform process, but as the result of intentional design choices, particularly during the fine-tuning stage, about how developers want their AI to behave. Those choices should have repercussions under tort law.

Picture first an AI system with the following alignment profile: the system (1) has been placed fully under human control; (2) is oriented exclusively toward its immediate user; (3) correctly divines user objectives and values; and (4) defers extra-legal ethical questions to user discretion.

---

200. *See, e.g.*, Ketan Ramakrishnan, *Tort Law Is the Best Way to Regulate AI*, WALL ST. J. (Sep. 24, 2024, at 14:30 ET), https://www.wsj.com/opinion/tort-law-is-the-best-way-to-regulate-ai-california-legal-liability-065e1220 [https://perma.cc/B7ZH-SRSM].

201. *See, e.g.*, KETAN RAMAKRISHNAN, GREGORY SMITH & CONOR DOWNEY, RAND, U.S. TORT LIABILITY FOR LARGE-SCALE ARTIFICIAL INTELLIGENCE DAMAGES (2024), https://www.rand.org/pubs/research_reports/RRA3084-1.html [https://perma.cc/3U-JU-NS5T].

202. This line of inquiry is most developed in the self-driving car context. *See, e.g.*, Jack Boeglin, *The Costs of Self-Driving Cars: Reconciling Freedom and Privacy with Tort Liability in Autonomous Vehicle Regulation*, 17 YALE J.L. & TECH. 171 (2015).

203. I hope to return to this subject at greater length in future work.

If such a system causes harm to a third-party, we might think that the user should be held responsible. After all, tort law already assigns vicarious liability to superiors for certain harms caused by their underlings (e.g., the doctrine of *respondeat superior*)[204] and the immediate user of such a system might also be the "least-cost avoider."[205] By contrast, we might think a developer should be held responsible for (and might be the least-cost avoider of) an AI system that: (1) has not been fully brought under control; (2) remains oriented in significant part to its developer; (3) inconsistently divines user objectives and values; and (4) imposes a thick layer of extra-legal ethical judgment on top of user choice.

In other words, how an AI system has been aligned—and the ways in which that alignment strategy tracks (or fails to track) the prescriptions of legal alignment—might prove quite relevant to assigning legal responsibility for the harms it causes. And that in turn means that the legal system should keep a close eye on the decisions developers make during the alignment-focused fine-tuning process.[206]

E. *Computational Law*

Legal alignment also offers a new perspective on the potential stakes and proper objectives of computational law—"the branch of legal informatics concerned with the automation of legal reasoning."[207] Existing forays into computational law have focused on goals like automating legal compliance, legal services, and government administration.[208]

This Article suggests a different direction for computational law—helping AI to become better aligned. Making legal concepts understandable to computer systems might prove useful not just for providing *us* legal advice and solving *our* legal problems; it might also help AI to assess the propriety of its *own* conduct.

For advances in computational law to best further AI alignment, a shift in focus is likely necessary. Current benchmarks for assessing an AI system's legal understanding often look to whether it can correctly identify the legal consequences of externally described circumstances (e.g., if *X* does *Y*, does that violate *Z* law?).[209] But legal alignment requires more

---

204. *See* RESTATEMENT (SECOND) OF TORTS § 317 (A.L.I. 1965).

205. For the idea that tort liability might be placed on the "least-cost avoider," see GUIDO CALABRESI, THE COST OF ACCIDENTS: A LEGAL AND ECONOMIC ANALYSIS (1970).

206. This proposal dovetails with Paul Ohm's suggestion that "scholars, policymakers, regulators, and judges [should] focus on the fine-tuning stage as the best stage to address many (but not all) problems with generative AI." Ohm, *supra* note 70, at 216.

207. For a brief overview of the field, see Michael R. Genesereth, *What Is Computational Law?*, SLS BLOGS: CODEX (Mar. 10, 2021), https://law.stanford.edu/2021/03/10/what-is-computational-law [https://perma.cc/WGE5-USYB].

208. For a compelling vision of how AI might improve legal understanding, legal culture, and legal justice, see ABDI AIDID & BENJAMIN ALARIE, THE LEGAL SINGULARITY: HOW ARTIFICIAL INTELLIGENCE CAN MAKE LAW RADICALLY BETTER (2023).

209. *See* NEEL GUHA, JULIAN NYARKO, DANIEL E. HO, CHRISTOPHER RÉ, ADAM CHILTON, ADITYA NARAYANA, ALEX CHOHLAS-WOOD, AUSTIN PETERS, BRANDON WALDON, DANIEL N.

than the ability to answer legal hypotheticals; it calls for an assessment of AI's own circumstances. This might require AI first to learn to reliably describe its own conduct and place in the world from an external perspective—a task that might require a greater degree of self-awareness than AI systems currently appear to possess.

Legal alignment thus throws down a gauntlet to computational law, as well as offers yet another reason to think that the field holds great promise: it could help to solve the alignment problem.

F. *Soft Law*

Legal alignment also helps inform "soft law" proposals to govern AI. Soft law undoubtedly has a role to play even if legal alignment is taken up—most obviously, because "hard" law is subject to constitutional constraints. Among other things, this means that law usually must be content-neutral (e.g., indifferent to whether a viewpoint on vaccine safety and efficacy is supported by science). Even if such constitutional constraints are warranted with respect to law promulgated by the government, it does not follow that they should be carried over *in toto* to soft law meant to influence privately developed and deployed AI systems.[210]

That said, many soft law proposals appear to echo standards of conduct that legally aligned AI might be expected to abide by anyway. A 2020 report from the Berkman Klein Center for Internet & Society identified the eight most common themes emphasized in AI soft law proposals: (1) privacy, (2) accountability, (3) safety and security, (4) transparency and explainability, (5) fairness and non-discrimination, (6) human control of technology, (7) professional responsibility, and (8) promotion of human values.[211] These eight principles likely capture *some* behavioral norms not supplied by legal alignment.

But consider again these common themes when listed side by side with the norms that a legally aligned artificial fiduciary would be obliged to follow. Such an AI system would have to observe: (1) rules of confidentiality (privacy); (2) a duty to account (accountability), (3) an obligation to act only for the principal's benefit (safety and security), (4) mandatory

---

ROCKMORE, DIEGO ZAMBRANO, DMITRY TALISMAN, ENAM HOQUE, FAIZ SURANI, FRANK FAGAN, GALIT SARFATY, GREGORY M. DICKINSON, HAGGAI PORAT, JASON HEGLAND, JESSICA WU, JOE NUDELL, JOEL NIKHAUS, JOHN NAY, JONATHAN H. CHOI, KEVIN TOBIA, MARGARET HAGAN, MEGAN MA, MICHAEL LIVERMORE, NIKON RASUMOV-RAHE, NILS HOLZENBERGER, NOAM KOLT, PETER HENDERSON, SEAN REHAAG, SHARAD GOEL, SHANG GAO, SPENCER WILLIAMS, SUNNY GANDHI, TOM ZUR, VARUN IYER & ZEHUA LI, LEGALBENCH: A COLLABORATIVELY BUILT BENCHMARK FOR MEASURING LEGAL REASONING IN LARGE LANGUAGE MODELS (2023).

210. Madeline Lamo & Ryan Calo, *Regulating Bot Speech*, 66 UCLA L. REV. 988, 988 (2019) (discussing "how efforts to regulate bots might run afoul of the First Amendment").

211. JESSICA FJELD, NELE ACHTEN, HANNAH HILLIGOSS, ADAM CHRISTOPHER NAGY & MADHULIKA SRIKUMAR, PRINCIPLED ARTIFICIAL INTELLIGENCE: MAPPING CONSENSUS IN ETHICAL AND RIGHTS-BASED APPROACHES TO PRINCIPLES FOR AI 4–5 (2020), https://dash.harvard.edu/server/api/core/bitstreams/c8d686a8-49e8-4128-969c-cb4a5f2ee145/content [https://perma.cc/U8QV-4B27].

disclosures (transparency and explainability), (5) non-discrimination law (fairness and non-discrimination), (6) duties of loyalty and obedience (human control of technology), (7) applicable standards of care (professional responsibility), and (8) an obligation to follow the law (promotion of human values). While not all these legal concepts match up perfectly with existing soft law proposals, the obvious overlap suggests that legal alignment has the potential to significantly narrow the work soft law measures need to do. In turn, soft law could focus on the subset of alignment questions that legal norms drawn from "hard" law alone cannot answer.

CONCLUSION

In offering a robust frame for tackling the concerns underlying the alignment problem, legal alignment represents an untapped resource for aligning AI to human objectives and values. But the onus is not just on alignment practitioners. It falls equally on lawmakers and legal scholars to proactively shape the legal norms to which AI should be aligned. Understanding how much the law has to say about alignment helps us understand how imperative it is that the law say the right things. The law is not fixed—and in the case of the nascent legal regime governing AI, it is bound to change. Further work is needed to drill down on how particular legal doctrines, theories, and modes of interpretation can best be repurposed as alignment solutions, as well as to further explore the implications of legal alignment for other schools of thought on AI governance. It is past time we stopped viewing law as part of the alignment problem and started viewing it as part of its solution.