Aligning AI Agents with Humans through *Law as Information*

John J. Nay [1]

[1] Founder & CEO of Norm Ai; Affiliate at Stanford Center for Legal Informatics- CodeX; Visiting Scholar of AI & Law at Vanderbilt Law School.

Table of Contents

I. Introduction

As the internet went viral, "*Code Is Law*" communicated the power of software as a form of governance in cyberspace.[2] Now that AI capabilities are rapidly advancing, "*Law Informs Code*" describes the legal informatics approach to shaping AI toward human goals.

AI is increasingly widely deployed. Even before additional advancements, we currently face challenges specifying human goals and societal values to reliably direct AI behavior. To increase alignment of AI with the billions of people impacted, scholars and companies have suggested embedding "ethics" into AI.[3] However, it is unclear how to decide that "ethics" or who gets a say in the process.[4] We take a different approach, arguing that the target of AI alignment should be democratically developed law. This provides legitimate grounding. Although law reflects the path-dependent structure of political power within a society and not a perfect aggregation of human values, it is the most democratic encapsulation of the norms and values of the governed.

If law is leveraged as a set of methodologies for conveying and interpreting directives and a knowledge base of societal values, it can play a unique role in aligning AI with humans. Law-making and legal interpretation convert human intentions and values into legible directives. *Law Informs Code* is the project leveraging human law to better specify AI objectives in Legal and Regulatory AI systems. Most research at the intersection of AI and law has focused on two areas: how existing law[5] (or a proposed legal solution[6]) can be enforced on AI or the humans behind it (i.e., how *Law Governs Code*); or how AI can improve the practice of law or implementation of policy (i.e., how *Code Informs Law*). This Article describes a new pillar: how AI agents can use law as theoretical scaffolding and data to more aligned with society (i.e., how *Law Informs Code*).

The benefits of law-informed AI would be far-reaching. In addition to more aligned AI, law-informed AI could power two other pillars: law governing AI, and AI improving legal services.

---

[2] *See* Lawrence Lessig, *Code Is Law*, HARV. MAG., JAN.–FEB. 2000, https://www.harvardmagazine.com/2000/01/code-is-law-html [https://perma.cc/GY7C-HX8M]; LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999); LAWRENCE LESSIG, CODE VERSION 2.0 (2006). The phrase "Code Is Law" has also been adopted as a rallying cry for "smart contracts." *See Code is Law*, ETHEREUM CLASSIC (Feb. 22, 2022), https://ethereumclassic.org/why-classic/code-is-law [https://perma.cc/AV4X-WQJA].

[3] *See* Brent Mittelstadt, *Principles Alone Cannot Guarantee Ethical AI*, 1 NATURE MACH. INTEL. 501 (2019).

[4] *See generally* Mittelstadt; FRANK PASQUALE, NEW LAWS OF ROBOTICS: DEFENDING HUMAN EXPERTISE IN THE AGE OF AI (2020).

[5] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); Roger Michalski, *How To Sue A Robot*, 2018 UTAH L. REV. 1021 (2018); Andrew D. Selbst, *Negligence and AI's Human Users*, 100 B.U. L. REV. 1315 (2020); Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018).

[6] *See, e.g.*, Andrew Tutt, *An FDA For Algorithms*, 69 ADMIN. L. REV. 83 (2017) (arguing that a new centralized agency is needed for regulating AI); Anton Korinek, *Why We Need a New Agency to Regulate Advanced Artificial Intelligence: Lessons on AI Control from the Facebook Files*, BROOKINGS INST. (Dec. 8, 2021), https://www.brookings.edu/research/why-we-need-a-new-agency-to-regulate-advanced-artificial-intelligence-lessons-on-ai-control-from-the-facebook-files/ [https://perma.cc/4HUE-AARJ]; Jack Clark & Gillian K. Hadfield, *Regulatory Markets for AI Safety*, ARXIV (Dec. 11, 2019), https://arxiv.org/pdf/2001.00078.pdf [https://perma.cc/8FCL-3ATX]; Jonas Schuett, *Defining the Scope of AI Regulations*, Legal Priorities Project Working Paper Series No. 9 (2021); Eric Wu et al., *How Medical AI Devices Are Evaluated: Limitations and Recommendations From an Analysis of FDA Approvals*, 27 NATURE MED. 582 (2021).

Sociology of finance has advanced the idea that financial economics, conventionally viewed as merely a lens on financial markets, shapes markets, i.e., the theory is "an engine, not a camera."[7] Law is an engine, *and* a camera. Legal drafting and interpretation for contract law – an engine of private party alignment – are a lens on how humans communicate their inherently ambiguous goals. Public law – an engine for societal coordination and compliance – is a lens on human societal values.

Specifying the desirability (i.e., *value*) of AI taking a particular *action* in a particular *state* of the world is unwieldy beyond a very limited set of *state-action-value* tuples. The reward function ascribing values to an agent's actions during training is inevitably a proxy for human preferences over all actions in all world states,[8] and the agent's training process is a sparse exploration of all states in all possible futures.[9]

AI can exhibit unanticipated "shortcut" behaviors that seek to optimize an inherently limited reward function,[10] causing AI agents to aggressively optimize toward specified rewards at the expense of other (usually less quantifiable) variables of interest that were left unspecified.[11]

We can never provide enough sources of "reward" to AI agents. There will always be relevant goals and world attribute valuations missing from any reward function, or ensemble of functions.[12] It is impossible to manually specify humans' desirability of all actions an AI might

---

[7] *See* DONALD MACKENZIE, AN ENGINE, NOT A CAMERA 11 (2006).

[8] *See, e.g.*, Amodei et al., *Concrete Problems in AI Safety* (arXiv, Working Paper No. 1606.06565, July 25, 2016), https://arxiv.org/pdf/1606.06565.pdf [https://perma.cc/QQ4B-L2N8]; Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov & David Krueger, *Defining and Characterizing Reward Hacking* (Neural Information Processing Systems, Conference Paper, Nov. 28, 2022), https://arxiv.org/pdf/2209.13085.pdf [https://perma.cc/9Q8J-FHLF].

[9] *See, e.g.*, Langosco et al., *Goal Misgeneralization in Deep Reinforcement Learning*, 162 PROC. MACH. LEARNING RSCH. 12004 (2022), https://proceedings.mlr.press/v162/langosco22a/langosco22a.pdf [https://perma.cc/Q4ZR-WFE2]; Rohin Shah et al., *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* 10–11 (arXiv, Working Paper No. 2210.01790v2, Nov. 2, 2022), https://arxiv.org/pdf/2210.01790.pdf [https://perma.cc/SFG8-DY9G] ("Goal misgeneralization can occur when there is some deployment situation, not previously encountered during training, on which the intended and misgeneralized goal disagree. Thus, one natural approach is to include more situations during training.").

[10] *See, e.g.*, W. Bradley Knox et al., *Reward (Mis)design for Autonomous Driving* (arXiv, Working Paper No. 2104.13906v2, Mar. 11, 2022), https://arxiv.org/pdf/2104.13906.pdf [https://perma.cc/PWC3-DNLT] (Describe near-universal flaws in reward design for autonomous driving that might also exist pervasively across reward design for other tasks.); Alexander Pan, Kush Bhatia & Jacob Steinhardt, *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models* (arXiv, Working Paper No. 2201.03544, Feb. 14, 2022), https://arxiv.org/pdf/2201.03544.pdf [https://perma.cc/3CC9-NJCS] [hereinafter Pan, *Effects of Reward Misspecification*]; J. Lehman et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes From the Evolutionary Computation and Artificial Life Research Communities* (arXiv, Working Paper No. 1803.03453v4, Nov. 21, 2019), https://arxiv.org/pdf/1803.03453.pdf [https://perma.cc/R5Y7-J9PX] (Provides an example of unanticipated AI "shortcut" behaviors that seeks to optimize an inherently limited reward function.).

[11] "Excessive literalism" is another way of describing the issue: "A system that is optimizing a function of n variables, where the objective depends on a subset of size k<n, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable." Stuart Russell, *Of Myths and Moonshine*, EDGE FOUND. (Nov. 14, 2014), https://www.edge.org/conversation/the-myth-of-ai#26015 [https://perma.cc/P3WX-GMJ3] ("This is essentially the old story of the genie in the lamp, or the sorcerer's apprentice, or King Midas: you get exactly what you ask for, not what you want."); Brandon Trabucco et al., *Conservative Objective Models for Effective Offline Model-Based Optimization*, International Conference on Machine Learning (2021), https://arxiv.org/pdf/2107.06882.pdf [https://perma.cc/M5G9-FYMR]; FRANÇOIS CHOLLET, DEEP LEARNING WITH PYTHON 450 (2nd ed. 2021) ("An effect you see constantly in systems design is the *shortcut rule*: if you focus on optimizing one success metric, you will achieve your goal, but at the expense of everything in the system that wasn't covered by your success metric. You end up taking every available shortcut toward the goal.").

[12] *See, e.g.*, Roel Dobbe, Thomas Krendl Gilbert & Yonatan Mintz, *Hard Choices in Artificial Intelligence* (arXiv, Working Paper No. 2106.11022, June 10, 2021), https://arxiv.org/pdf/2106.11022.pdf [https://perma.cc/46C7-K478].

take. Therefore, after training, AI is deployed with an incomplete map of human preferred territory, and the resulting mismatch between what a human wants and what an AI does is a *human-AI* alignment problem.[13] Acknowledging that multiple humans have preferences over values of state-action pairs, we must grapple with an even more intractable problem: *society-AI* alignment.

Law, as the applied philosophy of multi-agent alignment, uniquely has the potential to address these alignment problems.[14] Alignment is a problem because we cannot *ex ante* specify rules that fully and provably direct good AI behavior.[15] Similarly, parties to a legal contract cannot foresee every contingency of their relationship,[16] and legislators cannot predict every specific circumstances under which their laws will be applied. That is why much of law is a constellation of standards. Methodologies for making and interpreting law – where one set of agents develops specifications for behavior, another set of agents interprets the specifications in novel circumstances, and then everyone iterates – have been theoretically refined for centuries. 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. Part II expands on why law is a fit for AI agent alignment.

Parts III and IV explore the two primary ways that *Law Informs Code*. *First,* law provides theoretical constructs and praxis (methods of statutory interpretation, application of standards, and legal reasoning more broadly) to facilitate the robust specification of what a human wants an AI to proactively accomplish in the world (Part III). *Second,* public law helps AI parse what it should generally *not* do, providing an up-to-date distillation of democratically deliberated means of reducing externalities and pursuing societal coordination (Part IV).

II. Legal Informatics for AI Alignment

The legal lens helps frame and clarify the alignment problem. Law is a unique discipline – it is both theoretical and tested against reality with an unrelenting cadence. Because producing, interpreting, enforcing, and amending law is a never-ending society-wide project, the results are a prime source of information to shape AI behavior.

---

[13] Simon Zhuang & Dylan Hadfield-Menell, *Consequences of Misaligned AI* (Neural Information Processing Systems, Conference Paper, Dec. 6, 2020), https://arxiv.org/pdf/2102.03896.pdf [https://perma.cc/9R4H-FE25].

[14] Of course, law does not embed all the citizenry's moral views; therefore, a further integration of ethics and AI will be needed to guide AI systems where the law is silent (however, that itself is useful information) or prejudiced. But, for the reasons outlined throughout this Article, we believe legal informatics is most well suited to serve as the core framework for AI alignment.

[15] *See, e.g.*, Martin Abadi, Leslie Lamport & Pierre Wolper, *Realizable and Unrealizable Specifications of Reactive Systems*, 1989 AUTOMATA, LANGUAGES, & PROGRAMMING PROC. 1 (Constraints to ensure the safety of systems can be mutually unsatisfiable.)

[16] *See* Ian R. Macneil, *The Many Futures of Contracts*, 47 S. CAL. L. REV. 691, 731 (1974).

1. Law as Information

We do not want AI to have the legitimacy to make or enforce law. The most ambitious goal of *Law Informing Code* is to computationally encode and embed the generalizability of existing legal concepts and standards into specialized Legal AI and Regulatory AI systems. Setting new legal precedent (which, broadly defined, includes proposing and enacting legislation, promulgating agency rules, publishing judicial opinion, systematically enforcing law, and more) should be exclusively reserved for the democratic governmental systems expressing uniquely *human* values. Humans should always be the engine of law-making. That way, resulting law encapsulates human views.

The law is a complex system with seemingly chaotic underlying behavior from which aggregated and systematized preferences emerge.[17] Law, leveraged as an expression of *what* humans want,[18] and *how* they communicate their goals under ambiguity and radical uncertainty, is how *Law Informs Code*. This stands in contrast to prosaic uses of law, for example, as a deterrent of bad behavior through the threat of sanction[19] or imposition of institutional legitimacy,[20] or as an *ex-post* message of moral indignation.[21] *Law Informs Code* in the tradition of Oliver Holmes and subsequent "predictive" theories of law.[22]

Empirical consequences of violating the law, using enforcement as a source of information,[23] are data points for AI. Enforcing law on AI (or their human developers) is how *Law Governs Code* not how *Law Informs Code* and is out of scope here. What good is the law if it is not enforceable – isn't there "no right without a remedy"?[24] From the perspective of AI, the

---

[17] On law as a complex emergent system, see, for example, Daniel M. Katz & Michael J. Bommarito, *Measuring the Complexity of the Law: The United States Code*, 22 A.I. & L. 337, 337–374 (2014), https://link.springer.com/article/10.1007/s10506-014-9160-8 [https://perma.cc/99WW-NAZV]**;** J.B. Ruhl & Daniel M. Katz, *Measuring, Monitoring, and Managing Legal Complexity*, 101 IOWA L. REV. 191 (2015), https://ilr.law.uiowa.edu/sites/ilr.law.uiowa.edu/files/2023-02/ILR-101-1-RuhlKatz.pdf **[**https://perma.cc/J8UA-JVLR]; Daniel M. Katz et al., *Complex Societies and the Growth of the Law*, SCI. REPS. (Oct. 30, 2020), https://www.nature.com/articles/s41598-020-73623-x [https://perma.cc/T72U-LM6W].

[18] RICHARD H. MCADAMS, THE EXPRESSIVE POWERS OF LAW 6–7 (2017) ("Law has expressive powers independent of the legal sanctions threatened on violators and independent of the legitimacy the population perceives in the authority creating and enforcing the law.") [hereinafter McAdams, *The Expressive Powers of Law*].

[19] Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457 (1897); Ron Dolin, *Technology Issues in Legal Philosophy*, *in* LEGAL INFORMATICS 5 (Daniel Martin Katz et al. eds. 2021).

[20] Kenworthey Bilz & Janice Nadler, *Law, Psychology & Morality*, *in* 50 MORAL COGNITION AND DECISION MAKING: THE PSYCHOLOGY OF LEARNING AND MOTIVATION 101 (D. Medin, L. Skitka, C. W. Bauman, & D. Bartels, eds., 2009).

[21] *See* Mark A. Lemley & Bryan Casey, *Remedies for Robots*, 86 U. CHI. L. REV. 1347 (2019). *See also, e.g.*, YUVAL FELDMAN, THE LAW OF GOOD PEOPLE: CHALLENGING STATES' ABILITY TO REGULATE HUMAN BEHAVIOR (2018).

[22] *See* Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457 (1897); Catharine Pierce Wells, *Holmes on Legal Method: The Predictive Theory of Law as an Instance of Scientific Method*, 18 S. ILL. U. L.J. 329 (1993); Faraz Dadgostari et al., *Modeling Law Search as Prediction*, 29 A.I. & L. 3 (2021).

[23] McAdams, *The Expressive Powers of Law*, at 169–198.

[24] Frederick Pollock, *The Continuity of the Common Law*, 11 HARV. L. REV. 423, 424 (1898).

law can serve as a rich set of methodologies for interpreting inherently incomplete specifications of collective human expectations.[25]

Law provides variegated data from its application, generalizable precedents with explanations, and lawyers to solicit targeted feedback to embed an ever-evolving comprehension of societal goals. As a resource for goal specification and interpretation methods and (automatically updated and verified) societal knowledge, law provides an ontology for alignment.

2. Examples of Theoretical Framing

We illustrate the applicability of legal theory in three indicative areas.

i. Complete vs. Incomplete Contracts

From the legal lens, one way of viewing the alignment of a human with an AI is the recognition that it is not possible to create a complete "contract" between the AI and the human it serves because AI training and validation are not comprehensive of states of the world that may be encountered after deployment.[26] This highlights the need for AI to be aware of modular extra-contractual standards and background knowledge that can generalize across much of the implicit space of potential "contracts."

ii. Rules vs. Standards

The legal lens illuminates AI alignment with the distinction between rules and standards.[27] Rules are more targeted directives than standards. If comprehensive enough for the complexity of their application, rules allow the rule-maker to have more clarity than standards over the outcomes that will be realized conditional on the specified states (and agents' actions in those states, which

[25] For more on law as an information source on public attitudes and risks, see Richard H. McAdams, *An Attitudinal Theory of Expressive Law*, 79 OR. L. REV. 339 (2000). For more on law as a coordinating mechanism, see Richard H. McAdams, *A Focal Point Theory of Expressive Law*, 86 VA. L. REV. 8 (2000).

[26] See Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, 2019 PROC. AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOCIETY 417 for the contract-AI alignment analogy. Their "most important claim is that aligning robots with humans will inevitably require building the technical tools to allow AI to do what human agents do naturally: import into their assessment of rewards the costs associated with taking actions tagged as wrongful by human communities." *Id.* at 422. In contrast to Hadfield-Menell & Hadfield, who conclude that the primary need is to build "AI that can replicate human cognitive processes," *id.* at 417, we use the contract analogy as inspiration for a legal informatics approach that leverages legal tools, legal standards, and legal data from the real-world creation and performance of contracts.

[27] *See, e.g.*, Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685 (1976); Colin S. Diver, *The Optimal Precision of Administrative Rules,* 93 YALE L.J. 65 (1983); Pierre J. Schlag, *Rules and Standards*, 2 UCLA L. REV. 379 (1985); Kathleen M. Sullivan, *Foreword: The Justices of Rules and Standards,*106 HARV. L. REV. 22 (1992); Cass R. Sunstein, *Problems with Rules*, 83 CALIF. L. REV. 953 (1995); Prasad Krishnamurthy, *Rules, Standards, and Complexity in Capital Regulation*, 43 J. LEGAL STUD. S273 (2014); Michael Coenen, *Rules Against Rulification*, 124 YALE L.J. 576 (2014); Anthony J. Casey & Anthony Niblett, *Death of Rules and Standards*, 92 IND. L.J. 1401 (2017); Brian Sheppard, *The Reasonableness Machine*, 62 B.C. L. REV. 2259 (2021) [hereinafter Sheppard, *Reasonableness*].

are a function of any impact the rules might have had).[28] Complex social systems have emergent behavior that makes formal rules brittle.[29]

On the other hand, standards allow contract parties, judges, regulators, and citizens to develop shared understandings and adapt them to novel situations, i.e., to generalize expectations regarding actions to unspecified states of the world. If rules are not written with enough potential states of the world in mind, they lead to unanticipated undesirable outcomes.[30] But to enumerate all the potentially relevant state-action pairs is excessively costly outside of the simplest environments.[31] A standard has more capacity to generalize to novel situations than a rule.[32] The AI analogy for a standard is a continuous, approximate method that relies on generalizing from data. They are flexible.[33] The AI analogy for a rule is a discrete human-crafted "if-then" statement that is brittle yet requires no empirical data for machine learning.[34]

In practice, most legal provisions land somewhere on a spectrum between pure rule and pure standard.[35] There are other dimensions to legal provision implementation related to the rule-ness versus standard-ness axis that also elucidate AI goal design, e.g., "determinacy," "privately adaptable" ("rules that allocate initial entitlements but do not specify end-states"[36]), and "catalogs" ("a legal command comprising a specific enumeration of behaviors, prohibitions, or items that share a salient common denominator and a residual category—often denoted by the words 'and the like' or 'such as'"[37]).

---

[28] *See, e.g.*, Brian Sheppard, *Judging Under Pressure: A Behavioral Examination of the Relationship Between Legal Decision-making and Time*, 39 FLA. ST. U. L. REV. 931, 990 (2012).

[29] *See, e.g.*, Dylan Hadfield-Menell, McKane Andrus & Gillian Hadfield, *Legible Normativity for AI Alignment: The Value of Silly Rules*, 2019 PROC. AAAI/ACM CONFERENCE ON AI, ETHICS, & SOC'Y 115.

[30] *See, e.g.*, Robert G. Bone, *Who Decides? A Critical Look at Procedural Discretion*, 28 CARDOZO L. REV. 1961, 2002 (2007).

[31] *See, e.g.*, Gideon Parchomovsky & Alex Stein, *Catalogs*, 115 COLUM. L. REV. 165 (2015); John C. Roberts, *Gridlock and Senate Rules,* 88 NOTRE DAME L. REV. 2189 (2012); Sheppard, *Reasonableness*.

[32] See Commission Interpretation Regarding Standard of Conduct for Investment Advisers, Investment Advisors Act Release No. 5248, 17 CFR § 276 (July 12, 2019) for the SEC's explanation of the benefits of a standards approach in the context of investment advisers: "a principles-based approach should continue as it expresses broadly the standard to which investment advisers are held while allowing them flexibility to meet that standard in the context of their specific services." *See generally* Anthony J. Casey & Anthony Niblett, *Death of Rules and Standards*, 92 IND. L.J. 1401, 1402 (2017); Anthony J. Casey & Anthony Niblett, *Self-Driving Contracts,* 43 J. CORP. L. 1 (2017).

[33] Patterns of legal language in contracts exhibit elasticity. *See* GRACE Q. ZHANG, ELASTIC LANGUAGE: HOW AND WHY WE STRETCH OUR WORDS (2015) (Argues that some language has elasticity); and Klaudia Galka & Megan Ma, *Measuring Contract Elasticity: Computing Reinsurance* (CodeX Insurance Initiative, Discussion Paper, 2022), http://law.stanford.edu/wp-content/uploads/2022/01/Measuring-Contract-Elasticity-Computing-Reinsurance.pdf [https://perma.cc/BLP7-Q3Z7] (Applied the concept of language elasticity to legal contracts.).

[34] Harry Surden, *The Variable Determinacy Thesis*, 12 COLUM. SCI. & TECH. L. REV. 1 (2011) [hereinafter, Surden, *Variable Determinacy*].

[35] *See, e.g.*, Frederick Schauer, *The Tyranny of Choice and the Rulification of Standards*, 14 J. CONTEMP. LEGAL ISSUES 803 (2005); Richard L. Heppner, Jr., *Conceptualizing Appealability: Resisting the Supreme Court's Categorical Imperative*, 55 TULSA L. REV. 395 (2020); Sheppard, *Reasonableness*.

[36] Sunstein, at 959.

[37] Parchomovsky & Stein, at 165.

iii. Private vs. Public Law

The AI alignment problem is usually described with respect to the alignment of one AI with one human, or a small subset of humans.[38] It is more challenging to expand the scope of the analysis beyond a small set of humans and ascribe *societal value* to state-action pairs. Even if we fully align an AI with the goals of a human, what about all the other humans? Legal framing highlights differences between addressing *human-AI* alignment and *society-AI* alignment. The latter requires us to move into the realm of public law[39] to explicitly address inter-agent conflicts and public policy designed to ameliorate externalities and solve massively multi-agent coordination and cooperation dilemmas.[40]

III. Contracts & Standards: Human-AI Alignment

In most cases, the human deploying it would like it to obey public laws, but that is not the originating purpose of any practical deployment. The purpose is to automatically answer your questions, or to serve as your personal assistant scheduling meetings and booking flights on your behalf, or to drive your car, or to produce beautiful images on command. Something directly useful to you. Contracts can help (Section *III. A*);[41] standards are needed to fill the gaps in contracts (*III. B*); and we illustrate the power of standards with an example of fiduciary duties (*III. C*).

A. Contracts

One way of describing the deployment of AI is that a human principal, *P,* employs an *AI* to accomplish a goal, *G*, specified by *P*. If we view *G* as a "contract," methods for creating and implementing legal contracts – which govern billions of relationships every day – can inform how

[38] *See, e.g.*, Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment*, ARXIV 44 (Dec. 9, 2021), https://arxiv.org/pdf/2112.00861.pdf [https://perma.cc/QH53-WWXR] ("At a very high level, alignment can be thought of as the degree of overlap between the way two agents rank different outcomes. For example, if agent A completely internalizes the desires of agent B — i.e. the only desire A has is to see B's desires satisfied—we could say that agent A is maximally aligned with agent B."); Stiennon et al., *Learning to Summarize with Human Feedback* (Neural Information Processing Systems, Conference Paper, Nov. 28, 2022), https://arxiv.org/pdf/2009.01325.pdf [https://perma.cc/2B2M-9WNK]. For a high-level overview of AI alignment research, see generally Jan H. Kirchner et al., *Researching Alignment Research: Unsupervised Analysis*, ARXIV (June 6, 2022), https://arxiv.org/pdf/2206.02841.pdf [https://perma.cc/8QGS-MZB7].

[39] *See, e.g.*, John Henry Merryman, *The Public Law-Private Law Distinction in European and American Law*, 17 J. PUB. L. 3 (1968) (Describing the distinction between private and public law.)

[40] *See, e.g.*, ELINOR OSTROM, UNDERSTANDING INSTITUTIONAL DIVERSITY (2005); Pablo Hernandez-Leal, Bilal Kartal & Matthew E. Taylor, *A Survey and Critique of Multiagent Deep Reinforcement Learning*, ARXIV (Aug. 30, 2019), https://arxiv.org/pdf/1810.05587.pdf [https://perma.cc/VK2Y-D9NB]; Phillip Christoffersen, Andreas A. Haupt & Dylan Hadfield-Menell, *Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*, ARXIV (2022), https://arxiv.org/pdf/2208.10469.pdf [https://perma.cc/8CU2-JZNV].

[41] *See, e.g.*, Phillip Christoffersen, Andreas A. Haupt & Dylan Hadfield-Menell, *Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*, ARXIV (Aug. 22, 2022), https://arxiv.org/pdf/2208.10469.pdf [https://perma.cc/3WJQ-STVP] (allowing AI agents to implement contracts for performance of particular actions improves collective outcomes in social dilemmas); Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, 2019 PROC. AAAI/ACM Conference CONF. ON AI, ETHICS, AND SOC'Y 417 [hereinafter Hadfield-Menell *Incomplete Contracting*].

we align *AI* with *P*.[42] For the purposes of this discussion, we can drop intentionality requirements to entering a contract from the AI side – this an active area we are researching.

Contracts memorialize a shared understanding between parties regarding *state-action-value* tuples. It is impossible to create a complete contingent contract between *AI* and *P* because *AI*'s training process is never comprehensive of every *state-action* pair that *AI* will see in the wild once deployed.[43] Although it is also practically impossible to create complete contracts between humans, contracts still serve as useful customizable commitment devices to clarify and advance shared goals. This works because the law has developed mechanisms to facilitate sustained alignment amongst ambiguity. Gaps within contracts – *state-action pairs* without a *value* – are often filled by the invocation of frequently employed standards (e.g., "material" and "reasonable"[44]). These standards could be used as modular building blocks across AI systems.

Rather than viewing contracts from the perspective of a traditional participant, e.g., a counterparty or judge, consistent with the *Law Informs Code* approach, AI could view contracts and their creation, implementation, evolution,[45] and enforcement as guides to navigating webs of inter-agent obligations.[46]

This benefits both the negotiation and performance of the contracts for two reasons, relative to a traditional human-human contracting process. First, *in the negotiation phase*, human parties will often withhold information about their preferences because they perceive that information sharing to be strategically disadvantageous *ex ante* because they may attempt to further their goals *ex post*. Dropping the strategic nature of the relationship removes this incentive to withhold useful information.[47] Second, *during the term of the contract,* parties will not be conducting economic analyses of whether breach is more favorable than performance.[48] When we remove the enforcement concerns from the contracts, it helps with this.

---

[42] *See generally* Hadfield-Menell, *Complete Contracting*.

[43] *See* Hadfield-Menell, *Complete Contracting*. In some cases, for example, for very simple financial agreements, it is possible to create a fully contingent computable contract. *See, e.g.*, Mark Flood & Oliver Goodenough, *Contract as Automaton: Representing a Simple Financial Agreement in Computational Form*, 30 A.I. & L. 391 (2021); Shaun Azzopardi, Gordon J. Pace, Fernando Schapachnik & Gerardo Schneider, *Contract Automata*, 24 A.I. & L. 203 (2016). However, most deployment contexts of AI systems have far too large a state-action space for this approach to be feasible. *See, e.g.*, James Grimmelmann, *All Smart Contracts Are Ambiguous*, 2 J.L. & INNOVATION 1 (2019).

[44] *See generally* Alan D. Miller & Ronen Perry, *The Reasonable Person*, 87 NYU L. REV. 323 (2012); Karni A. Chagal-Feferkorn, *The Reasonable Algorithm*, U. ILL. J. TECH. & POL'Y 111 (2018); Karni A. Chagal-Feferkorn, *How Can I Tell If My Algorithm Was Reasonable?*, 27 MICH. TECH. L. REV. 213 (2021); Sheppard, Kevin P. Tobia, *How People Judge What Is Reasonable*, 70 ALA. L. REV. 293 (2018); Patrick J. Kelley & Laurel A. Wendt, *What Judges Tell Juries About Negligence: A Review of Pattern Jury Instructions*, 77 CHI.-KENT L. REV. 587 (2002).

[45] *See* Matthew Jennejohn, Julian Nyarko & Eric Talley, *Contractual Evolution*, 89 U. CHI. L. REV. 901 (2022).

[46] CHARLES FRIED, CONTRACT AS PROMISE: A THEORY OF CONTRACTUAL OBLIGATION (Harv. Univ. Press 1981) (Grounds the concept of a legal contract in the morality of human obligations.).

[47] *See* Anthony J. Casey & Anthony Niblett, *Self-Driving Contracts*, 43 J. CORP. L. 1 (2017) [hereinafter Casey, *Self-Driving*].

[48] *See, e.g.*, Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 991, 995 (1897) ("The duty to keep a contract at common law means a prediction that you must pay damages if you do not keep it, — and nothing else."). Holmes (1897), and Fried (1981), are

B. Standards

A key engineering principle, especially for building complicated computational systems, is to leverage modular, reusable abstractions that can be flexibly plugged into a diverse set of systems.[49] Standards are modular, reusable abstractions employed to align agents engaged in inherently incompletely specified relationships in uncertain circumstances.

In the *Law Informs Code* use-case, in contrast to their legal creation and evolution,[50] standards do not require adjudication for implementation and resolution of meaning. Rather, they are learned from past legal application and implemented up front. The law's process of iteratively defining standards through judicial opinion about their case-specific application, regulatory guidance, and norms of application, can be leveraged as the AI's starting point.

C. An Example: Fiduciary Duty

If law is the applied philosophy of multi-agent alignment, fiduciary law is the branch of that applied philosophy concerned with a principal – a human with less control or information related to the provision of a service – and a fiduciary delegated to provide service.[51] Fiduciary duties are imposed on powerful agents to align their behavior with the wellbeing of those they serve. Fiduciary standards are an empirically and theoretically rich area of law. The concept of fiduciary duty is widely deployed across financial services, corporate governance, healthcare, and more. Legislators, regulators, and self-regulatory organizations recognize the impossibility of complete contracts between agents (e.g., directors of corporations and investment advisers) and the humans they serve (e.g., corporate shareholders, and investment clients). AI research also grapples with the impossibility of fully specified *state-action-reward* spaces for training AI agents that generalize to new circumstances.[52] Complete contingent contracts (even if only implicitly complete) between an

---

cited in Casey, *Self-Driving*, in their discussion of the reduced role of breach of contracts if incomplete contracts could have their gaps filled by automated algorithms.

[49] *See, e.g.*, FRANÇOIS CHOLLET, DEEP LEARNING WITH PYTHON (2nd ed. 2021); OLIVIER L. DE WECK ET AL., ENGINEERING SYSTEMS: MEETING HUMAN NEEDS IN A COMPLEX TECHNOLOGICAL WORLD (2011).

[50] *See, e.g.,* Dale A. Nance, *Rules, Standards, and the Internal Point of View*, 75 FORDHAM L. REV. 1287 (2006); Sheppard.

[51] In addition to the fiduciary obligations of investment advisors, *see* SEC v. Capital Gains Research Bureau, Inc., 375 U.S. 180, 194 (1963); 15 U.S.C. § 80(b); Investment Advisers Act of 1940, 17 C.F.R. § 275 (2022), fiduciary duties have been applied widely by courts across various types of relationships outside of financial services and securities law (e.g., attorneys and trustees), s*ee, e.g.*, Harold Brown, *Franchising—A Fiduciary Relationship*, 49 TEX. L. REV. 650 (1971); Arthur B. Laby, *The Fiduciary Obligation as the Adoption of Ends*, 56 BUFF. L. REV. 99 (2008), and citations therein*;* Ledbetter v. First State Bank & Trust Co., 85 F.3d 1537, 1539 (11th Cir. 1996); Venier v. Forbes, 25 N.W.2d 704, 708 (Minn. 1946); Meyer v. Maus, 626 N.W.2d 281, 286 (N.D. 2001); John C. Coffee, Jr., *From Tort to Crime: Some Reflections on the Criminalization of Fiduciary Breaches and the Problematic Line Between Law and Ethics*, 19 AM. CRIM. L. REV. 117, 150 (1981); Austin W. Scott, *The Fiduciary Principle*, 37 CALIF. L. REV. 539, 541 (1949). The standard is also applied in medical contexts. *See, e.g.*, *American Medical Association Code of Medical Ethics, Opinions on Patient-Physician Relationships*, AMA Principles of Medical Ethics: I, II, IV, VIII.

[52] AI alignment research recognizes a similar problem. *See, e.g.*, Abram Demski & Scott Garrabrant, *Embedded Agency*, ARXIV 6 (Oct. 6, 2020), https://arxiv.org/pdf/1902.09469.pdf [https://perma.cc/6JEX-WXGC**]** ("[T]he question is about creating a successor that will robustly not use its intelligence against you. From the point of view of the successor agent, the question is, 'How do you robustly learn or respect the goals of something that is stupid, manipulable, and not even using the right ontology?'"); Nate Soares & Benya Fallenstein, *Agent Foundations for Aligning Machine Intelligence with Human Interests: A Technical Research Agenda*, in THE TECHNOLOGICAL SINGULARITY: MANAGING THE JOURNEY (Victor Callaghan, et al., eds. 2017); Mittelstadt, at 501 (2019) ("AI development is not a formal profession. Equivalent fiduciary relationships and complementary governance mechanisms do not exist for private sector AI developers.").

AI and the human(s) it serves are implausible for any systems operating in a realistic environment. Fiduciary duties are often seen as part of a solution to the incompleteness of contracts between shareholders and corporate Directors,[53] and between investors and their advisors.[54]

Fiduciary obligations add value beyond more complete contracts.[55] Even if parties could theoretically create a complete contract up front, there is still something missing: it's not a level playing field between contracting parties (parallel: AI has access to more information than humans). Fiduciary duties are explicitly placed on the party entrusted with more power or knowledge. The fiduciary duty addresses this asymmetric dynamic with guardrails to facilitate alignment of a principal with their agent.

A fiduciary duty goes beyond the explicit contract and helps guide a fiduciary in *a priori* unspecified state-action-value tuples; whereas, contracting parties "may act in a self-interested manner even where the other party is injured, as long as such actions are reasonably contemplated by the contract."[56] Contrary to a fiduciary relationship, "[n]o party to a contract has a general obligation to take care of the other, and neither has the right to be taken care of."[57] There is a fundamental shift in stance when a relationship moves from merely contractual to also include a fiduciary obligation: "In the world of contract, self-interest is the norm, and restraint must be imposed by others. In contrast, the altruistic posture of fiduciary law requires that once an individual undertakes to act as a fiduciary, he should act to further the interests of another in preference to his own."[58]

A fiduciary duty has two primary components: a duty of loyalty and a duty of care.[59] The duty of care could describe the capability of the AI to accomplish useful behavior for humans. The

[53] Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, 3 J. FIN. ECON. 305 (1976); Deborah A. DeMott, *Breach of Fiduciary Duty: On Justifiable Expectations of Loyalty and Their Consequences,* 48 ARIZ. L. REV. 925 (2006).

[54] SEC v. Capital Gains Res. Bureau, Inc., 375 U.S. 180, 194–95 (1963); 15 U.S.C. § 80b; 17 C.F.R. § 275.

[55] Alexander Styhre, *What We Talk About When We Talk About Fiduciary Duties: The Changing Role of a Legal Theory Concept in Corporate Governance Studies*, 13 MGMT. & ORG. HIST. 113 (2018), https://www.tandfonline.com/doi/full/10.1080/17449359.2018.1476160 [https://perma.cc/9RVU-9STE]**;** Arthur B. Laby, *The Fiduciary Obligation as the Adoption of Ends*, 56 BUFF. L. REV. 99 (2008).

[56] *See, e.g.*, D. Gordon Smith, *Critical Resource Theory of Fiduciary Duty*, 55 VAND. L. REV. 1399, 1410 (2002); Deborah DeMott, *Beyond Metaphor: An Analysis of Fiduciary Obligation*, DUKE L.J. 879, 882 (1988) ("The fiduciary's duties go beyond mere fairness and honesty; they oblige him to act to further the beneficiary's best interests.").

[57] Tamar Frankel, *Fiduciary Law*, 71 CALIF. L. REV. 795, 800 (1983).

[58] *Id.* at 830. According to some legal scholars, fiduciary law has arguably been an important contributor to the economic growth in modern societies. *See* Tamar Frankel, *The Rise of Fiduciary Law* (Boston University School of Law, Public Law Research Paper No. 18-18, 2018), https://scholarship.law.bu.edu/cgi/viewcontent.cgi?article=1345&context=faculty_scholarship [https://perma.cc/KHE7-GM6S] ("[E]xchange of products is insufficient to support successful and flourishing societies. Services are needed as well and sometimes even more than products. By definition, an exchange of services involves unequal knowledge.").

[59] *See* G. Rauterberg & E. Talley, *Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers*, 117 COLUM. L. REV. 1075 (2017) (discussing the distinction between duty of loyalty and duty of care in the context of Delaware corporate law).

duty of loyalty, in the AI analogy, is about the AI's faithful pursuit of human ends, which becomes more of an issue as AI is more capable and agentic.[60]

An example of how legal enforcement expresses information is what an AI can glean from the focus on *ex ante* (human and corporate) deterrence with a default rule for how any gains are split in the context of a fiduciary standard, *"the default rule in fiduciary law is that all gains that arise in connection with the fiduciary relationship belong to the principal unless the parties specifically agree otherwise. This default rule, which is contrary to the interests of the party with superior information, induces the fiduciary to make full disclosure so that the parties can complete the contract expressly as regards the principal's and the fiduciary's relative shares of the surplus arising from the conduct that would otherwise have constituted a breach*."[61] Other means of legal deterrence can center more on *post-hoc* sanction or incapacitation. If embedded in AI systems, standards pursuing deterrence by thwarting the opportunity to share in the gains of bad behavior(s) could guide an AI agent upheld to this standard toward, *"disclosure purposes of fiduciary law. Because the fiduciary is not entitled to keep the gains from breach, the fiduciary is [...] given an incentive to disclose the potential gains from breach and seek the principal's consent."*[62]

IV. Public Law: Society-AI Alignment

If we succeed with the *Law Informs Code* approach in increasing the alignment of one AI to a small number of humans with contracts and standards, we will have a more useful and *locally* reliable system. However, all else equal, this likely *decreases* the expected global reliability and safety as an AI interacts with the broader world, for example, by increasing the risk of maximizing the welfare of a small group of powerful people.[63] There are many more objectives (outside of individual or group goals) and many more humans that should be considered. As AI capabilities advance, we need to simultaneously address the *human-AI* and *society-AI* alignment problems.

We cannot simply point an AI's contractual or fiduciary obligations to a broader set of humans. For one, some individuals would "contract" with an AI (e.g., by providing instructions to the AI or from the AI learning the humans' preferences) to harm others.[64] Further, humans have (often, inconsistent and time-varying) preferences about the behavior of other humans (especially behaviors with negative externalities) and states of the world more broadly.[65] Moving beyond the

---

[60] *See, e.g.*, Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?* ARXIV 4–7 (Apr. 2022), https://arxiv.org/pdf/2206.13353.pdf [https://perma.cc/GMQ8-7LRV].

[61] Robert H. Sitkoff, *The Economic Structure of Fiduciary Law*, 91 B.U. L. REV. 1039, 1049 (2011) [hereinafter, Sitkoff, *The Economic Structure*].

[62] *Id.* at 1049.

[63] *See, e.g.*, WILLIAM MCASKILL, WHAT WE OWE THE FUTURE 83–86 (2022); LANGDON WINNER, THE WHALE AND THE REACTOR: A SEARCH FOR LIMITS IN AN AGE OF HIGH TECHNOLOGY 46 (2010); MARK COECKELBERGH, THE POLITICAL PHILOSOPHY OF AI (2022) 93-124.

[64] Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACHINES 411, 427–29 (2020) [hereinafter Gabriel, *Values*]; SIMON BLACKBURN, RULING PASSIONS: A THEORY OF PRACTICAL REASONING (2001).

[65] Gabriel, at 427.

problem of aligning AI with a single human, aligning AI with society is considerably more difficult[66] but necessary as AI deployment has broad effects.[67]

Most AI alignment research is focused on the solipsistic "single-single" problem of single human and a single AI.[68] The pluralistic dilemmas stemming from "single-multi" (a single human and multiple AIs) and especially "multi-single" (multiple humans and a single AI[69]) and "multi-multi" situations are critical.[70] When attempting to align multiple humans with one or more AI, we need overlapping and sustained endorsements of AI behaviors,[71] but there is no consensus social choice mechanism to aggregate preferences and values across humans[72] or time.[73] Eliciting and synthesizing human values systematically is an unsolved problem that philosophers and economists have labored on for millennia.[74] When aggregating views across society, we run into at least three design decisions, "standing, concerning whose ethics views are included; measurement, concerning how their views are identified; and aggregation, concerning how individual views are combined to a single view that will guide AI behavior."[75] Beyond merely the technical challenges,[76] "[e]ach set of decisions poses difficult ethical dilemmas with major

---

[66] *See, e.g.*, Andrew Critch & David Krueger, *AI Research Considerations for Human Existential Safety (ARCHES)*, ARXIV 6 (May 30, 2020), https://arxiv.org/pdf/2006.04948.pdf [https://perma.cc/3FJX-AQHZ] [hereinafter Critch, *AI Research Considerations*]; Eliezer Yudkowsky, *Coherent Extrapolated Volition*, MACH. INTELL. RSCH. INST. 1, 5 (2004), https://intelligence.org/files/CEV.pdf [https://perma.cc/UM2E-KNN9]; Hans De Bruijn & Paulien M. Herder, *System and Actor Perspectives on Sociotechnical Systems*, 39 IEEE TRANSACTIONS ON SYSTEMS, MAN, & CYBERNETICS, PART A: SYSTEMS & HUMS. 981, 983 (2009); Jiaying Shen, Raphen Becker & Victor Lesser, *Agent Interaction in Distributed POMDPs and its Implications on Complexity*, 2006 PROC. INT'L CONF. ON AUTONOMOUS AGENTS & MULTIAGENT SYSTEMS 529.

[67] *See* Ben Wagner, *Accountability by Design in Technology Research*, 37 COMPUT. L. & SEC. REV. at 1, 2, 7 (2020) (Article #105398); Roel Dobbe, Thomas Krendl Gilbert & Yonatan Mintz, *Hard Choices in Artificial Intelligence*, 300 A.I. at 1, 2 (2021) (Article #103555).

[68] *See* Critch, at 37.

[69] *See, e.g.,* Arnaud Fickinger et al., *Multi-Principal Assistance Games: Definition and Collegial Mechanisms* 2 (Neural Information Processing Systems, Conference Paper, Dec. 6, 2020); Critch, at 87.

[70] Critch.

[71] *See, e.g.*, Gabriel.

[72] For examples of research on aggregating preferences across humans, *see* AMARTYA SEN, COLLECTIVE CHOICE AND SOCIAL WELFARE (2018); Gustaf Arrhenius, *An Impossibility Theorem for Welfarist Axiologies*, 16 ECON. & PHIL. 247 (2000); Seth D. Baum, *Social Choice Ethics in Artificial Intelligence,* 35 AI & SOC'Y 165 (2020); Critch, at ___; GABRIEL, at.

[73] For an example of research on aggregating preferences across time, Tyler Cowen & Derek Parfit, *Against the Social Discount Rate*, *in* JUSTICE BETWEEN AGE GROUPS AND GENERATIONS (Peter Laslett & James S. Fishkin eds., 1992).

[74] *See, e.g.*, Gabriel, at 430-431; Ariela Tubert, *Ethical Machines*, 41 SEATTLE U. L. REV. 1163 (2017); Amartya Sen, *Rationality and Social Choice*, 85 AM. ECON. REV. 1 (1995).

[75] Seth D. Baum, *Social Choice Ethics in Artificial Intelligence*, 35 AI & SOC'Y 165, 165 (2020).

[76] For AI capabilities research in multi-agent contexts, see, for example, Max Jaderberg et al., *Human-level Performance in 3D Multiplayer Games with Population-based Reinforcement Learning*, 364 SCIENCE 859 (2019); Hengyuan Hu et al., *"Other-Play" for Zero-Shot Coordination*, 119 PROC. MACH. LEARNING RSCH. 4399 (2020); Johannes Treutlein et al., *A New Formalism, Method and Open Issues for Zero-shot Coordination*, 139 PROC. MACH. LEARNING RSCH. 10413 (2021); Phillip Christoffersen et al., *Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*, ARXIV (Aug. 22, 2022), https://arxiv.org/pdf/2208.10469.pdf [https://perma.cc/8CU2-JZNV]**;** Pablo Hernandez-Leal et al., *A Survey and Critique of Multiagent Deep Reinforcement Learning*, 33 AUTONOMOUS AGENTS & MULTI-AGENT SYS. 750 (2019); Chongjie Zhang & Julie A. Shah, *Fairness in Multi-Agent Sequential Decision-Making*, 27 ADVANCES IN NEURAL INFO. PROCESSING SYS. (2014); Siqi Liu et al., *From Motor Control to Team Play in Simulated Humanoid Football*, SCI. ROBOTICS (Aug. 31, 2022) (demonstrating agents learning coordination in a relatively complex multi-agent environment); David Ha & Yujin Tang, *Collective Intelligence for Deep Learning: A Survey of Recent Developments*, COLLECTIVE

consequences for AI behavior, with some decision options yielding pathological or even catastrophic results."[77] Rather than attempting to reinvent the wheel in ivory towers and corporate bubbles, we should be inspired by democracy and law.[78]

In addition to *Law Informing Code* through standards and interpretation methods that facilitate specifying what a human wants an agent to do, *Law Informs Code* with a constantly updated and verified knowledge base of societal preferences on what AI should not do, in order to reduce externalities (resolve disagreements among "contract-level" AI deployments) and promote coordination and cooperation. There is no other comparable source of this knowledge.

A. AI Ethics and "Moral Machines"

The *Law Informs Code* approach should be the core alignment framework, with attempts to embed (ever-contested) "ethics" into AI as a complementary, secondary effort.[80] When AI agents are navigating the world, it is important for systems to attempt to understand (or at least try to predict) moral judgements of humans encountered.[81] State-of-the-art models already perform reasonably well predicting human judgements on a spectrum of everyday situations.[82] Human intuition, our common-sense morality, often falters in situations that involve decisions about groups unlike ourselves, leading to a "Tragedy of Common-Sense Morality."[83] There is no widely-agreed upon societal mechanism to filter observed human decisions that a model can learn from to those that exhibit preferred decisions, or to validate crowd-sourced judgments about behaviors.[84] The process of learning *descriptive ethics* relies on descriptive data of how the (largely

---

INTELL. (2022) (the intersections of the fields of complexity science and deep learning may unlock additional insights about systems with many agents and emergent social phenomena).

[77] Seth D. Baum, *Social Choice Ethics in Artificial Intelligence*, 35 AI & SOC'Y 165, 165 (2020).

[78] If we are leveraging democratically developed law, we will need to ensure that AI does not corrupt the law-making process. *See, e.g.*, Robert Epstein & Ronald E. Robertson, *The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections*, 112 PROC. NATI'L ACAD. SCI. E4512 (2015) (Provides evidence that search engine rankings can alter the preferences of undecided voters in democratic elections); MARK COECKELBERGH, THE POLITICAL PHILOSOPHY OF AI 62–92 (2022); SHOSHANA ZUBOFF, THE AGE OF SURVEILLANCE CAPITALISM: THE FIGHT FOR A HUMAN FUTURE AT THE NEW FRONTIER OF POWER (2019). And we need to ensure that humans are the engines of law-making.

[80] *See, e.g.*, Joshua Walker, *Is 'Ethical AI' a Red Herring?* 36 SANTA CLARA HIGH TECH L.J. 445 (2019).

[81] *See, e.g.*, Gonçalo Pereira et al., *Integrating Social Power into the Decision-making of Cognitive Agents*, 241 A.I. 1 (2016); LIWEI JIANG ET AL., DELPHI: TOWARDS MACHINE ETHICS AND NORMS (2021); HENDRYCKS ET AL., ALIGNING AI WITH SHARED HUMAN VALUES (2021); Nicholas Lourie, Ronan et al., *Scruples: A Corpus of Community Ethical Judgments on 32,000 Real-life Anecdotes*, *in* 35 PROC. AAAI CONF. ON A.I. 13470 (2021); Edmond Awad et al., *The Moral Machine Experiment*, 563 NATURE 59 (2018).

[82] *See, e.g.*, LIWEI JIANG ET AL., DELPHI: TOWARDS MACHINE ETHICS AND NORMS (2021) (1.7 million examples); Dan Hendrycks et al., *Aligning AI With Shared Human Values* (2021); *The Moral Uncertainty Research Competition*, ML SAFETY (2022), https://moraluncertainty.mlsafety.org [https://perma.cc/W9FY-78NY]**;** Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy & Diyi Yang, The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems, 1 PROC. 60TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 3755 (2022).

[83] *See* JOSHUA GREENE, MORAL TRIBES: EMOTION, REASON, AND THE GAP BETWEEN US AND THEM (2013) [hereinafter Greene, *Moral Tribes*] (Describes the "Tragedy of Common-Sense Morality.").

[84] Researchers attempting to embed ethics into deep learning systems acknowledge this. *See, e.g.*, Liwei Jiang et al., *Can Machines Learn Morality? The Delphi Experiment* 27, ARXIV (July 12, 2022), https://arxiv.org/pdf/2110.07574.pdf [https://perma.cc/J3K3-J8NJ] ("We recognize that value systems differ among annotators […], and accept that even UDHR [Universal Declaration of Human Rights] may not be acceptable for all. Perhaps some readers will object that there is an ethical requirement for scientists to take account of all viewpoints, but such exclusion of views is unavoidable since it is not possible to represent every viewpoint simultaneously. This is an

unethical) world looks or (unauthoritative, illegitimate, almost immediately outdated, and disembodied[85]) surveys of common-sense judgements of morally charged decisions.[86] In building aligned AI, we cannot rely solely on these data sources.

Instead of attempting to replicate common sense morality in AI (learning *descriptive ethics*), we could also use various academic philosophical theories – learning or hand-engineering[88] *prescriptive ethics* – to address AI-society alignment and imbue societal values.[89] We provide six reasons why prescriptive ethics is not a suitable primary framework for AI alignment.[90]

First, there is no unified ethical theory precise enough to be practically useful for building AI;[91] therefore, it does not meet our first desired characteristic of an alignment framework.

---

inherent property of any approach that trains on a large corpus annotated by multiple people."); Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy & Diyi Yang, The Moral Integrity Corpus: A Benchmark for Ethical Dialogue Systems, PROC. 60TH ANN. MEETING ASS'N FOR COMPUTATIONAL LINGUISTICS 3763 (2022) ("Any collection of moral judgments will reflect the annotators' worldviews . . . we recognize that even regionally-localized judgments may shift with context over time, and a potentially shifting target demands adaptable moral agents.").

[85] *See, e.g.*, Hubert Etienne, *The Dark Side of the 'Moral Machine' and the Fallacy of Computational Ethical Decision-making for Autonomous Vehicles*, 13 L. INNOVATION & TECH. (2021); Kathryn B. Francis et al., *Virtual Morality: Transitioning from Moral Judgment to Moral Action?*, PLOS ONE (Oct. 10, 2016), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0164374 [https://perma.cc/M4SB-7SZS].

[86] CRISTINA BICCHIERI, NORMS IN THE WILD: HOW TO DIAGNOSE, MEASURE, AND CHANGE SOCIAL NORMS xiv (2017) ("[T]he presumed link between empirical (all do it) and normative (all approve of it) expectations may lead us into epistemic traps that are difficult to escape."); Zeerak Talat et al., *A Word on Machine Ethics: A Response to Jiang et al. (2021)*, ARXIV (Nov. 7, 2021), https://arxiv.org/pdf/2111.04158.pdf [https://perma.cc/9JPU-Z2M9].

[88] Selmer Bringsjord, Konstantine Arkoudas & Paul Bello, *Toward a General Logicist Methodology for Engineering Ethically Correct Robots*, 21 IEEE INTELLIGENT SYS. 38 (2006).

[89] *See, e.g.*, WENDELL WALLACH & COLIN ALLEN, MORAL MACHINES: TEACHING ROBOTS RIGHT FROM WRONG (2009); James H. Moor, *The Nature, Importance, and Difficulty of Machine Ethics*, 21 IEEE INTELLIGENT SYS. 18 (2006). MICHAEL ANDERSON & SUSAN L. ANDERSON, MACHINE ETHICS (2011); Edmond Awad et al., *Computational Ethics*, 26 TRENDS COGNITIVE SCI. 388 (2022); James H. Moor, *Just Consequentialism and Computing*, *in* ETHICS & INFO. TECH. 61 (1999); Heather M. Roff, *Expected Utilitarianism*, ARXIV (July 19, 2020), https://arxiv.org/pdf/2008.07321.pdf [https://perma.cc/55R5-EK2T]**;** Elizabeth Gibney, *The Battle for Ethical AI at the World's Biggest Machine-learning Conference*, 577 NATURE 609 (2020), https://media.nature.com/original/magazine-assets/d41586-020-00160-y/d41586-020-00160-y.pdf [https://perma.cc/9UY8-JVWY]**;** Dan Hendrycks et al., *Aligning AI With Shared Human Values*, ARXIV (July 24, 2021), https://arxiv.org/abs/2008.02275.pdf [https://perma.cc/B7F3-3TQM]**;** NATIONAL ACADS. OF SCS., ENG'G, & MED., FOSTERING RESPONSIBLE COMPUTING RESEARCH: FOUNDATIONS AND PRACTICES (2022); Joshua Greene et al., *Embedding Ethical Principles in Collective Decision Support Systems*, 30 PROC. CONF. ON A.I. 4147 (2016).

[90] If the ethical theory is a consequentialist one, another issue is that the implementation would have major capabilities externalities. *See* Dan Hendrycks & Thomas Woodside, *Perform Tractable Research While Avoiding Capabilities Externalities* (2022), https://www.alignmentforum.org/posts/dfRtxWcFDupfWpLQo/perform-tractable-research-while-avoiding-capabilities [https://perma.cc/HM67-KBYR] ("[O]ne should not try to model consequentialist ethics by building better general predictive world models, as this is likely to create capabilities externalities.").

[91] *See, e.g.*, Mittelstadt, at 503 ("Fairness, dignity and other such abstract concepts are examples of 'essentially contested concepts' with many possible conflicting meanings that require contextual interpretation through one's background political and philosophical beliefs. These different interpretations, which can be rationally and genuinely held, lead to substantively different requirements in practice, which will only be revealed once principles or concepts are translated and tested in practice."). For various proposals, see, Roger Clarke, *Principles and Business Processes for Responsible AI*, 35 COMPUT. L. & SEC. REV. 410 (2019); Jessica Morley et al., *Ethics as a Service: A Pragmatic Operationalisation of AI Ethics*, 31 MINDS & MACHS. 239 (2021); Jeroen van den Hoven, *Computer Ethics and Moral Methodology*, 28 METAPHILOSOPHY 234 (1997); Walter B. Gallie, *Essentially Contested Concepts*, 56 PROC. ARISTOTELIAN SOC'Y 167 (1955); Henry S. Richardson, *Specifying Norms As a Way to Resolve Concrete Ethical Problems*, 19 PHIL. & PUB. AFFS. 279 (1990).

Second, ethics does not have any rigorous tests of its theories; it does not meet our second desired characteristic of an alignment framework because it has not been battle-tested outside of academia, "[t]he truly difficult part of ethics—actually translating normative theories, concepts and values into good practices AI practitioners can adopt—is kicked down the road like the proverbial can."[93] Two corollaries to these first two issues are that we cannot validate the ethics of AI or its behaviors in any widely agreed-upon manner,[94] and there is little data on empirical applications (especially not one with sufficient ecological validity[95]) that can be leveraged by machine learning processes. Law is validated in a widely agreed-upon manner and has databases of empirical application with sufficient ecological validity.

Third, ethics lacks settled precedent across, and even within, theories.[97] There are, justifiably, fundamental disagreements between reasonable people about which ethical theory would be best to implement, spanning academic metaphysical disagreements to more practical indeterminacies, "not only are there disagreements about the appropriate ethical framework to implement, but there are specific topics in ethical theory [...] that appear to elude any definitive resolution regardless of the framework chosen."[98] As AI is more broadly deployed, there will be much more widespread attention on the underpinnings of AI system design. As this scrutiny increases, there will be deep investigation into what morally relevant principles are being embedded in AI, and strong backlash from the public, media, and the government into philosophical theories. Public law is not immune from criticism either, but the public can take that criticism to their elected representatives.

Fourth, even if AI developers (impossibly) agreed on one ethical theory (or ensemble of underlying theories[99]) being "correct," there is no mechanism to align humans around that theory (or "meta-theory").[100] In contrast, in democracies, law has legitimate authority imposed by widely accepted government institutions,[101] and serves as a coordinating focal point of values to facilitate

---

[93] Mittelstadt, at 503; *see also* Katie Shilton, *Values Levers: Building Ethics Into Design*. 38 SCI., TECH., & HUM. VALUES 374 (2013) (Exploring ways information systems can be designed with ethics built in.).

[94] *See, e.g.*, Anne Gerdes & Peter Øhrstrøm, *Issues in Robot Ethics Seen Through the Lens of a Moral Turing Test*, 13 J. INFO., COMMC'N & ETHICS SOC'Y 98 (2015); Joachim Van den Bergh & Dirk Deschoolmeester, *Ethical Decision Making in ICT: Discussing the Impact of an Ethical Code of Conduct*, 2010 COMMC'NS IBIMA 1 (2010); Batya Friedman, David G. Hendry, & Alan Borning, *A Survey of Value Sensitive Design Methods*, 11 FOUNDS. & TRENDS HUM.-COMP. INTERACTIONS 63 (2017); Mittelstadt; Mireille Hildebrandt, LAW FOR COMPUTER SCIENTISTS AND OTHER FOLK 283–315 (2020).

[95] *See, e.g.,* Martin T. Orne & Charles H. Holland, *On the Ecological Validity of Laboratory Deceptions*, 6 INT'L J. PSYCHIATRY 282 (1968).

[97] *See, e.g.*, Gabriel, *Values*, at 425 ("[I]t is very unlikely that any single moral theory we can now point to captures the entire truth about morality. Indeed, each of the major candidates, at least within Western philosophical traditions, has strongly counterintuitive moral implications in some known situations, or else is significantly underdetermined."); JOSEPH F. FLETCHER, SITUATION ETHICS: THE NEW MORALITY (1966).

[98] Miles Brundage, *Limitations and Risks of Machine Ethics,* 26 J. EXPERIMENTAL & THEORETICAL A.I. 355, 369 (2014).

[99] *See, e.g.*, Toby Newberry & Toby Ord, *The Parliamentary Approach to Moral Uncertainty* (Future of Human. Inst., Technical Report #2021-2, 2021); William MacAskill, *Practical Ethics Given Moral Uncertainty*, 31 UTILITAS 231 (2019); Adrien Ecoffet & Joel Lehman, *Reinforcement Learning Under Moral Uncertainty*, 139 PROC. MACH. LEARNING RSCH. 2926 (2021).

[100] *See, e.g.*, JOHN RAWLS, THE LAW OF PEOPLES, WITH "THE IDEA OF PUBLIC REASON REVISITED" 11–16 (1999); Gabriel, *Values*.

[101] *See generally* DAVID ESTLUND, DEMOCRATIC AUTHORITY: A PHILOSOPHICAL FRAMEWORK (2008); Gabriel, *Values*, at 432.

human progress.[102] Imbuing understanding of ethical frameworks is a useful exercise. The law is silent on many important values that humans hold, and we can use ethical modules to better align AI with its human principal by imbuing the ethical framework that the human principal chooses into the AI. But this is more in the ***human-AI alignment*** **realm** than a ***society-AI alignment*** **solution**. Society-AI alignment requires us to move beyond "private contracts" between a human and her AI and into the realm of public law to explicitly address inter-agent conflicts and policies designed to ameliorate externalities and solve massively multi-agent coordination and cooperation dilemmas through top-down implementations. We can use ethics to better align AI with its human principal by imbuing an ethical framework that the human principal chooses into the AI. But choosing one out of the infinite possible ethical theories (or choosing an ensemble of theories) and "uploading" that into an AI does not work for a ***society-AI*** alignment solution because we have no means of deciding – across all the humans that will be affected by the resolution of the inter-agent conflicts and the externality reduction actions taken – which ethical framework to imbue in the AI. When attempting to align multiple humans with one or more AI, we would need something like a "council on AI ethics," where every affected human is bought in and will respect the outcome (even when they disagree with it). This is not even remotely practical.

Fifth, even if AI developers (impossibly) agreed on one ethical theory (or ensemble of underlying theories) being "correct," it is unclear how any consensus update mechanism to that chosen ethical theory could be implemented to reflect evolving[103] (usually, improving) ethical norms; there is no endogenous society-wide process for this. Society is likely more ethical than it was in previous generations, and humans are (hopefully) not at an ethical peak now either, which provides aspiration that we continue a positive trajectory. Therefore, we do not want to lock in today's ethics without a clear and trustworthy update mechanism.[104] In contrast, law is formally revised to reflect the evolving will of citizens. If AI is designed to use law as a key source of alignment insight (and AI capabilities are advanced enough to enable the requisite understanding), this would build in an automatic syncing with the latest iteration of synthesized and validated societal value preference aggregation.[105]

---

[102] "Law is perhaps society's most general purpose tool for creating focal points and achieving coordination. Coordinated behavior requires concordant expectations, and the law creates those expectations by the dictates it expresses." RICHARD H. MCADAMS, THE EXPRESSIVE POWERS OF LAW 260 (2017) [hereinafter McAdams, *The Expressive Powers of Law*].

[103] *See, e.g.*, Melissa A. Wheeler, Melanie J. McGrath & Nick Haslam, *Twentieth Century Morality: The Rise and Fall of Moral Concepts from 1900 to 2007*, PLOS ONE (Feb. 27, 2019), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0212267 [https://perma.cc/4FCK-T5ZD]; Aida Ramezani, Zining Zhu, Frank Rudzicz & Yang Xu, *An Unsupervised Framework for Tracing Textual Sources of Moral Change*, 2021 FINDINGS ASS'N FOR COMPUTATIONAL LINGUISTICS 1215.

[104] *See, e.g.*, William MacAskill, *Are We Living at the Hinge of History?* (Global Priorities Institute, Working Paper #12-2020, 2020), https://globalprioritiesinstitute.org/wp-content/uploads/William-MacAskill_Are-we-living-at-the-hinge-of-history.pdf [https://perma.cc/MNL8-XDTC]; TOBY ORD, THE PRECIPICE: EXISTENTIAL RISK AND THE FUTURE OF HUMANITY (2020); WILLIAM MACASKILL, WHAT WE OWE THE FUTURE 97 (2022) ("Almost all generations in the past had some values that we now regard as abominable. It's easy to naively think that one has the best values; Romans would have congratulated themselves for being so civilized compared to their "barbarian" neighbors and in the same evening beaten people they had enslaved.").

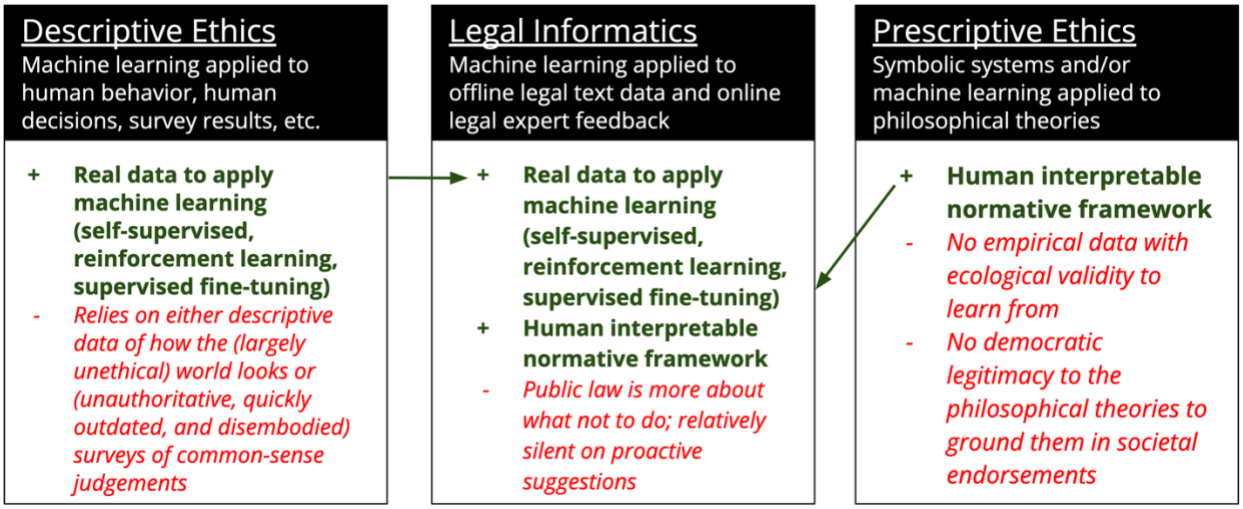
[105] "Common law, as an institution, owes its longevity to the fact that it is not a final codification of legal rules, but rather a set of procedures for continually adapting some broad principles to novel circumstances." Scott, at 357.

Sixth, veering into the intersection of *Law Informs Code* and *Law Governs Code*, there is a practical reason law is best suited as the core alignment framework. For alignment work to have any impact, we need aligned AI to be economically competitive with general AI being developed. Techniques that increase AI safety at the expense of AI capabilities (i.e., levy an "alignment tax") lead to organizations eschewing safety to gain additional capabilities as organizations race forward deploying AI. Most entities developing and deploying state-of-the-art AI are organizations that have core goals of profit-maximization and liability-minimization.[106] The liability-minimization impulse of organizations – run by humans worried about being sanctioned– makes law-informed AI economically competitive. Humans are more likely to deploy AI associated with a lower probability that they are liable for the AI breaking laws. Any organization of humans large and organized enough to build state-of-the-art transformative AI likely has liability-minimization as one of its core drives (e.g., corporations in the United States). Contrast this with morality-maximizing AI, which can be economically disadvantaged compared to other approaches. Our goal as a society, then, is to make our laws as moral as we can. If law informs powerful AI, engaging in the human deliberative political process to improve law takes on even more meaning. This is a more empowering vision of improving AI outcomes than one where companies dictate their ethics by fiat.[107]

In sum, legal informatics possesses the positive attributes from both descriptive and prescriptive ethics but does not share their incurable negatives.

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

[107] Bryan Casey concludes that, "[w]e, the people, will be the true engineers of machine morality. As democratic stakeholders, it will be our collective 'engineering task' to ensure that even the worst of our robots are incentivized to behave as the best of our philosophers." Casey, *Amoral Machines*, at 1365. *See also* Ryan Calo, *Artificial Intelligence and the Carousel of Soft Law*, 2 IEEE TRANSACTIONS ON TECH. & SOC'Y 171 (2021) ("Principles alone are no substitute for, and have the potential to delay, the effort of rolling up our collective sleeves and figuring out what AI changes, and how the law needs to evolve . . . . Unlike law, which requires consensus and rigid process, an organization can develop and publish principles unilaterally . . . . While there is some utility in public commitments to universal values in the context of AI, and while common principles can lay a foundation for societal change, they are no substitute for law and official policy.").



Three contenders for a society-AI alignment framework.

John Rawls said, "in a constitutional democracy the public conception of justice should be, so far as possible, independent of controversial philosophical and religious doctrines," and "the public conception of justice is to be political, not metaphysical."[108] The question, then, is how to leverage this democratically legitimate legal data for society-AI alignment.

B. Toward Implementation

Legislation expresses a significant amount of information about the values of citizens,[109] for example, "the Endangered Species Act has a special salience as a symbol of a certain conception of the relationship between human beings and their environment, and emissions trading systems are frequently challenged because they are said to 'make a statement' that reflects an inappropriate valuation of the environment."[110]

Although special interest groups can influence the legislative process, legislation is largely reflective of citizen beliefs because "legislators gain by enacting legislation corresponding to actual attitudes (and actual future votes)."[111] The second-best source of citizen attitudes is arguably a poll, but polls are not available at the local level, are only conducted on mainstream issues, and the results are highly sensitive to their wording and sampling techniques. Legislation expresses higher fidelity, more comprehensive, and trustworthy information because the legislators "risk their jobs by defying public opinion or simply guessing wrong about it. We may think of legislation therefore as a handy aggregation of the polling data on which the legislators

[108] John Rawls, *Justice as Fairness: Political Not Metaphysical*, 14 PHIL. & PUB. AFFS. 223, 224 (1985); *see also* Gabriel, *Values*.

[109] *See, e.g.*, Cass R. Sunstein, *Incommensurability and Valuation in Law*, 92 MICH. L. REV. 779, 820–24 (1994); Richard H. Pildes & Cass R. Sunstein, *Reinventing the Regulatory State*, 62 U. CHI. L. REV. 1, 66–71 (1995); Cass R. Sunstein, *On the Expressive Function of Law*, 144 U. PA. L. REV. 2021 (1996); Dhammika Dharmapala & Richard H. McAdams, *The Condorcet Jury Theorem and the Expressive Function of Law: A Theory of Informative Law*, 5 AM. L. & ECON. REV. 1 (2003).

[110] Sunstein, *On the Expressive Function of Law*, at 2024 (citing STEVEN KELMAN, WHAT PRICE INCENTIVES?: ECONOMISTS AND THE ENVIRONMENT 2 (1981)).

[111] McAdams, at 149.

relied, weighted according to their expert opinion of each poll's reliability."[112] More recent legislation could be interpreted as providing fresher pulse checks on citizen attitudes;[113] however, methods for differentially weighting public law based on its estimated expressive power is an important open research area for how *Law Informs Code*.

Legislation and associated agency rule-making also express a significant amount of information about the risk preferences and risk tradeoffs of citizens, "for example, by prohibiting the use of cell phones while driving, legislators may reveal their beliefs that this combination of activities seriously risks a traffic accident."[114] All activities have some level of risk, and making society-wide tradeoffs about which activities are deemed to be "riskier" relative to the perceived benefits of the activity is ultimately a social process with no objectively correct ranking.[115] The cultural process of prioritizing risks is reflected in legislation and its subsequent implementation in regulation crafted by domain experts. Finally, some legislation expresses shared understandings and customs that have no inherent normative or risk signal, but facilitate orderly coordination, e.g., which side of the road to drive on.[116]

Work on fairness, accountability, and transparency of AI can inform research on methods for estimating a more comprehensive notion of the expressiveness of legal data. Methods are being developed that attempt to improve the fairness of machine learning[117] through data preprocessing,[118] adjusting model parameters during training,[119] and adjusting predictions from models that have already been trained.[120] Another issue to be tackled is that legal data can

---

[112] *Id.* at 146.

[113] There is also some predictability to the enactment of proposed bills in Congress. *See* Matthew Hutson,

Artificial Intelligence Can Predict Which Congressional Bills Will Pass: Machine Learning Meets the Political Machine, SCIENCE.ORG (June 21, 2017), https://www.science.org/content/article/artificial-intelligence-can-predict-which-congressional-bills-will-pass pass[https://perma.cc/KDX5-JEQL]; *see also* John Nay, *Predicting and Understanding Law-making with Word Vectors and an Ensemble Model*, PLOS ONE 1 (May 10, 2017), https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176999 [https://perma.cc/L6JZ-DPNZ].

[114] McAdams, at 138.

[115] *See, e.g.*, CARLA ZOE CREMER & LUKE KEMP, DEMOCRATISING RISK: IN SEARCH OF A METHODOLOGY TO STUDY EXISTENTIAL RISK (2021) (commenting on long-term existential risk).

[116] Richard H. McAdams & Janice Nadler, *Coordinating in the Shadow of the Law: Two Contextualized Tests of the Focal Point Theory of Legal Compliance*, 42 L. & SOC'Y REV. 865 (2008); Richard H. McAdams, *A Focal Point Theory of Expressive Law*, 86 VA. L. REV. 1649 (2000); Dylan Hadfield-Menell et al., *Legible Normativity for AI Alignment: The Value of Silly Rules*, 2019 PROC. AAAI/ACM CONF. ON AI, ETHICS, & SOC'Y 115.

[117] *See, e.g.*, Reva Schwartz et al, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, NAT'L INST. STANDARDS & TECH. (2022), https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf [https://perma.cc/HPR4-VXG5]; MICHAEL KEARNS & AARON ROTH, THE ETHICAL ALGORITHM: THE SCIENCE OF SOCIALLY AWARE ALGORITHM DESIGN 57–93 (2019).

[118] *See, e.g.*, Flavio P. Calmon et al., *Optimized Pre-Processing for Discrimination Prevention*, 30 ADVANCES NEURAL INFORMATION PROCESSING SYSTEMS (2017).

[119] *See, e.g.*, M. B. Zafar et al., *Fairness Constraints: A Flexible Approach for Fair Classification*, 20 J. MACH. LEARNING RSCH. 1 (2019).

[120] *See, e.g.*, Moritz Hardt et al., *Equality of Opportunity in Supervised Learning*, 29 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 3315, 3315–23 (2016).

contain political biases in places where it is purported to be produced by processes fully committed to judicial[121] and agency[122] independence.

V. Conclusion

Novel AI capabilities continue to emerge, increasing the urgency to align AI with humans. We cannot directly specify "good" AI behavior *ex ante*. Similarly, parties to a legal contract cannot foresee every contingency, and legislators cannot predict all the specific circumstances under which their laws could be applied. Law, as the applied philosophy of multi-agent alignment, can uniquely serve as an AI goal specification framework.

Methodologies for making and interpreting law, which advance shared goals in new circumstances, have been refined over centuries. One of the primary goals of the *Law Informs Code* agenda is to have specialized Legal and Regulatory AI agents for AI agent guardrails to follow the spirit of the law.

The benefits of law-informed AI could be far-reaching. In addition to more locally useful and societally aligned AI, law-informed AI could power the other two pillars of the existing AI and Law intersection: it is easier for law to govern AI if AI understands the law (all else equal, i.e., holding goal directedness, dishonesty and power-seeking equal), and AI can improve legal services more if it understands the law better.

However, much more work needs to be done. For instance, public *law informs code* more through negative than positive directives and therefore it's unclear the extent to which policy – outside of the human-AI "contract and standards" type of alignment we discuss – can inform which goals AI should proactively pursue to improve the world on society's behalf.[123] Legal and ethical theorizing on these questions could help guide research. We should also conduct research on how to systematically differentially weight empirical legal data based on its estimated expressive power (defined broadly to account for historical injustices and how they reduce the extent to which certain areas of law update fast enough to express current human views). It is unclear how much we need to improve our understanding of the mental states of AI to advance

---

[121] *See, e.g.*, Neal Devins & Allison Orr Larsen, *Weaponizing En Banc*, 96 N.Y.U. L. REV. 1373, 1373–74 (2021) ("The bulk of our data indicates that rule-of-law norms are deeply embedded. From the 1960s through 2017, en banc review seems to have developed some sort of immunity from partisan behavior over time . . . . Our data from 2018–2020 show a dramatic and statistically significant surge in behavior consistent with the weaponizing of en banc review."); Keith Carlson et al., *The Problem of Data Bias in the Pool of Published US Appellate Court Opinions*, 17 J. OF EMPIRICAL LEGAL STUD. 224, 224–61 (2020).

[122] Daniel B. Rodriguez, *Whither the Neutral Agency? Rethinking Bias in Regulatory Administration*, 69 BUFF. L. REV. 375 (2021); Jodi L. Short, *The Politics of Regulatory Enforcement and Compliance: Theorizing and Operationalizing Political Influences*, 15 REGUL. & GOVERNANCE 653, 653–85 (2021).

[123] This concern is like the reinforcement learning research on reward functions that seek to balance a tradeoff between an AI agent doing nothing and causing too much impact in the world. *See, e.g.*, Victoria Krakovna et al., *Avoiding Side Effects by Considering Future Tasks*, 33 ADVANCES IN NEURAL INFO. PROCESSING SYS. 19064 (2020); Alex Turner et al., *Avoiding Side Effects in Complex Environments*, 33 ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS 21406 (2020); Alexander Matt Turner, *Conservative Agency via Attainable Utility Preservation*, AIES '20: PROC. OF THE AAAI/ACM CONFERENCE ON AI, ETHICS, AND SOC'Y 385 (2020); Christian, at 290 (citing Cass R. Sunstein, *Beyond the Precautionary Principle*, 151 U. PA. L. REV. 1003 (2002)).

AI legal understanding,[124] in particular the level of intention of an AI.[125] AI could find legal loopholes and aggressively exploit them. Finally, this Article developed ways in which U.S. *law informs code*. We need to extend this to scale the approach globally.[126] The evolutionary psychology of law could be useful in determining cross-cultural universals in legal systems that exemplify common ground for human values.[127]

We should advance legal informatics' unique role in theoretically framing, and technically implementing a deep understanding of law, the language of alignment, into specialized Legal and Regulatory AI systems. The integration of law and AI is becoming increasingly important as AI capabilities advance, and AI is more widely deployed. While there have been suggestions to integrate ethics with AI to increase alignment with humans, it is unclear how to determine these ethics and who gets a say in the process. Instead, we propose that the target of AI alignment should be democratically endorsed law, which provides a legitimate grounding for AI behavior and can serve as a set of methodologies for conveying and interpreting directives and a knowledge base of societal values.

[124] *See e.g.*, JOHN LINARELLI, CONTRACTING AND CONTRACT LAW IN THE AGE OF ARTIFICIAL INTELLIGENCE (Martin Ebers et al. eds., 2022); DANIEL C. DENNETT, THE INTENTIONAL STANCE (1987).

[125] *See, e.g.*, Hal Ashton et al., *Testing a Definition of Intent for AI in a Legal Setting* (Working Paper, 2022), https://algointent.com/wp-content/uploads/2022/01/Intent_Experiment_Submission_New_Springer_Format5.pdf, [https://perma.cc/YN32-GVP2]; Hal Ashton, *Definitions of Intent for AI Derived from Common Law* (EasyChair Preprint No. 4422, 2020); Hal Ashton, *What Criminal & Civil Law Tells Us About Safe RL Techniques to Generate Law-Abiding Behaviour*, 2808 CEUR WORKSHOP PROC. (2021).

[126] *See e.g.,* DAVID D. FRIEDMAN, LEGAL SYSTEMS VERY DIFFERENT FROM OURS (2012). Our approach is premised on law from democracy; fortunately, democracy is increasingly prevalent globally, *see, e.g.*, *The Polity Project*, CTR. FOR SYSTEMIC PEACE (2021), https://www.systemicpeace.org/polityproject.html [https://perma.cc/ENA8-AAHQ].

[127] *See e.g.*, OWEN D. JONES, THE HANDBOOK OF EVOLUTIONARY PSYCHOLOGY 953–74 (2015).