Forthcoming in *Northwestern Journal of Technology and Intellectual Property*, Volume 20

Law Informs Code: A Legal Informatics Approach to Aligning Artificial Intelligence with Humans

John Nay*
Center for Legal Informatics, Stanford University
September 13, 2022

ABSTRACT — Artificial Intelligence (AI) capabilities are rapidly advancing, and highly capable AI could cause radically different futures depending on how it is developed and deployed. We are currently unable to specify human goals and societal values in a way that reliably directs AI behavior. Specifying the desirability (*value*) of an AI system taking a particular *action* in a particular *state* of the world is unwieldy beyond a very limited set of *value-action-states*. The purpose of machine learning is to train on a subset of states and have the resulting agent generalize an ability to choose high value actions in unencountered circumstances. But the function ascribing values to an agent's actions during training is inevitably an incredibly incomplete encapsulation of the breadth of human values, and the training process is unavoidably a sparse exploration of states pertinent to all possible futures. Therefore, after training, AI is deployed with a coarse map of human preferred territory and will often choose actions unaligned with our preferred paths.

Law is a computational engine that converts opaque human values into legible and enforceable directives. *Law Informs Code* is the research agenda attempting to capture that complex computational process of human law, and embed it in AI. Similar to how parties to a legal contract cannot foresee every potential "if-then" contingency of their future relationship, and legislators cannot predict all the circumstances under which their proposed bills will be applied, we cannot *ex ante* specify "if-then" rules that provably direct good AI behavior. Legal theory and practice have developed arrays of tools to address these specification problems, a language of alignment. For instance, legal standards allow humans to develop shared understandings and adapt them to novel situations, i.e., to generalize expectations regarding actions taken to unspecified states of the world. In contrast to more prosaic uses of the law (e.g., as a deterrent of bad behavior through the threat of sanction), leveraged as an expression of *how* humans communicate their goals, and *what* society values, *Law Informs Code*.

We describe how the data generated by legal processes and the theoretical constructs and practices of law (methods of law-making, statutory interpretation, contract drafting, applications of standards, legal reasoning, etc.) can facilitate the robust specification of inherently vague human goals for AI. This helps with *human-AI* alignment and the local usefulness of AI. Toward *society-AI* alignment, we present a framework for understanding law as the applied philosophy of multi-agent alignment, harnessing public law as an up-to-date knowledge base of democratically endorsed values ascribed to state-action pairs. Although law is partly a reflection of historically contingent political power – and thus not a perfect aggregation of citizen preferences – if properly parsed, its distillation offers a legitimate computational comprehension of human goals and societal values. If law eventually informs powerful AI, engaging in the deliberative human political process to improve law takes on even more meaning.

* Contact: jnay@nyu.edu. This draft Article benefited from discussions with James Daily, Jonathan Gilligan, Megan Ma, and Oliver Schwartz. The research project at the Stanford Center for Legal Informatics this draft is part of can be found here: https://law.stanford.edu/projects/a-legal-informatics-approach-to-aligning-artificial-intelligence-with-humans. This Article represents my personal views and not those of Stanford, Brooklyn Investment Group, or any other organization. Nothing herein should be construed as investment or financial advice.

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.[1] As Artificial Intelligence (AI) capabilities advance[2] with new model architectures[3] scaled across massive data sets,[4] "*Law Informs Code*" could be the catchphrase for the legal informatics approach to shaping AI to be more in line with human values. AI is increasingly widely deployed.[5] More powerful AI technology could cause radically different

---

[1] *See,* Lawrence Lessig, *Code Is Law* (2000); Lawrence Lessig, *Code and Other Laws of Cyberspace* (1999); Lawrence Lessig, *Code 2.0* (2006). The phrase "Code Is Law" has also been adopted as a rallying cry for blockchain-based "smart contracts," *see, e.g., Code is Law*, Ethereum Classic https://ethereumclassic.org/why-classic/code-is-law.

[2] *See, e.g.,* Scott Reed et al., *A Generalist Agent* (2022) (A multi-modal, multi-task AI agent, "Gato," that can successfully play the Atari video game, caption images, chat, stack blocks with a robot, and more, all using the same neural network with the same parameters; "the recent progress in generalist models suggests that safety researchers, ethicists, and most importantly, the general public, should consider their risks and benefits."); Michael Ahn et al., *Do as I Can, Not as I Say: Grounding Language in Robotic Affordances*, arXiv:2204.01691 (2022) (An AI system that provides "real-world grounding by means of pretrained skills, which are used to constrain the model to propose natural language actions that are both feasible and contextually appropriate. The robot can act as the language model's "hands and eyes," while the language model supplies high-level semantic knowledge about the task. We show how low-level skills can be combined with large language models so that the language model provides high-level knowledge about the procedures for performing complex and temporally-extended instructions, while value functions associated with these skills provide the grounding necessary to connect this knowledge to a particular physical environment."); Holden Karnofsky, *AI Timelines: Where the Arguments, and the "Experts," Stand* (Sep 7, 2021) https://www.cold-takes.com/where-ai-forecasting-stands-today/ (Based on analyzing and synthesizing relevant technical reports forecasting AI capabilities from multiple approaches, Karnofsky derives an estimate that "there is more than a 10% chance we'll see transformative AI [powerful enough to bring us into a new, qualitatively different future] within 15 years (by 2036); a ~50% chance we'll see it within 40 years (by 2060); and a ~2/3 chance we'll see it this century (by 2100)"); Ajeya Cotra, *Two-year Update On My Personal AI Timelines*, LessWrong (August 2022) (In 2020, Cotra published an influential report forecasting a median estimate of transformative AI in 2050, and in August 2022, she shortened the estimate to 2040.) [Hereinafter Cotra, *Two-year Update On My Personal AI Timelines*]; Jacob Steinhardt, *AI Forecasting: One Year In* (2022) ("progress on ML benchmarks happened significantly **faster** than forecasters expected [...] Progress on a *robustness* benchmark was slower than expected, and was the only benchmark to fall short of forecaster predictions. This is somewhat worrying, as it suggests that machine learning capabilities are progressing quickly, while safety properties are progressing slowly." (Emphasis in original)); the most recent large survey of AI researchers (August 2022) on transformative AI timelines can be found at https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

[3] An example of an architecture enabling advanced capabilities is the Transformer (a model of sequential data with stacked self-attention layers and residual connections), outlined in the landmark paper, Vaswani et al., *Attention is All you Need* (2017). *See infra* Section II for more on the Transformer. The alternate title for this Article is "*Law Is All You Need"* (as Vaswani et al. do not seriously claim that attention is all you need for capable AI systems, *Law Is All You Need* should also not be interpreted as a literal claim about AI alignment).

[4] *See, e.g.,* Leo Gao et al., *The Pile: An 800GB Dataset of Diverse Text for Language Modeling* (2020).

[5] For a summary of AI deployment, *see, generally,* Daniel Zhang et al., *The AI Index 2022 Annual Report*, Stanford Institute for Human-Centered Artificial Intelligence (March 2022), https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf. For legal discussions of AI deployments, *see, e.g.,* Danielle K. Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 Washington Law Review 1-33 (2014); E. Joh, *The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing*, Harvard Law and Policy Review vol. 10, no. 15, 15-42 (2016); C. Muñoz, M. Smith & D. J. Patil, *Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights*, Executive Office of the U.S. President, Washington, D.C. (May 2016). On large language model deployments, *see, e.g.,* Matthew Hutson, *Robo-writers: The Rise and Risks of*

futures depending on the safety of its development and deployment.[6] A summer 2022 survey of hundreds of AI researchers estimated an aggregate forecast time of 37 years for a 50% chance of "high–level machine intelligence" ("when unaided machines can accomplish every task better and more cheaply than human workers").[7] Natural language processing (NLP) is a key sub-domain of AI for legal informatics, so surveys of NLP researchers are of particular interest. A separate summer 2022 survey of hundreds of NLP researchers found that 57% believe that "recent research has advanced us toward AGI [artificial general intelligence] in some significant way," and 73% "agree that labor automation from AI could plausibly lead to revolutionary societal change in this century, on at least the scale of the Industrial Revolution."[8] Even without additional technological advancements, we already face significant challenges specifying human goals and societal values in a way that reliably directs AI behavior at scale.[9]

Our thesis is that law, if leveraged as an information source, can play a unique role in theoretically framing and implementing the alignment of AI behavior with human goals. Law is a computational process that converts human values into legible[10] and enforceable directives. *Law Informs Code* is the research agenda attempting to capture that process of human law, and embed it in AI models. Most of the research at the intersection of the fields of AI and law has focused on two areas thus far: how existing law[11] (or a proposed legal solution[12]) applies to AI and the humans

---

Language-Generating AI, Nature 591.7848 22-25 (2021); Sam Manning et al., *A Research Agenda for Assessing the Economic Impacts of Code Generation Models* (2022) https://cdn.openai.com/papers/Economic_Impacts_Research_Agenda.pdf.

[6] For a longer-term framing of potential AI impacts, *see, e.g.*, Amanda Askell, *Ensuring the Safety of Artificial Intelligence*, in The Long View: Essays on policy, philanthropy, and the long-term future, edited by Natalie Cargill & Tyler John (2021); Henry Kissinger, Eric Schmidt & Daniel P. Huttenlocher, *The Age of AI: And Our Human Future* (2021); For a nearer-term framing of potential AI risks, see the recent work by the U.S. National Institute of Standards, *e.g., AI Risk Management Framework: Second Draft* (August 18, 2022) https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf.

[7] *See, 2022 Expert Survey on Progress in AI* (August 23, 2022) https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/. Other surveys of AI researchers also estimate non-trivial median probabilities that, in the next decade, "AI capabilities emerge that could radically transform welfare, wealth, or power, to an extent comparable to the nuclear revolution or even the industrial revolution. These possibilities are strikingly neglected, in part because they involve massive global and intergenerational externalities. There is thus a high leverage opportunity to address what may be the most important global issue of the 21st century." Allan Dafoe, *AI Governance: A Research Agenda* (Centre for the Governance of AI, Future of Humanity Institute, University of Oxford, 2018) at 5.

[8] Julian Michael et al., *What Do NLP Researchers Believe? Results of the NLP Community Metasurvey*, (2022) https://arxiv.org/abs/2208.12852 at 11 (36% of respondents believe "it is plausible that AI could produce catastrophic outcomes in this century, on the level of all-out nuclear war.").

[9] *See, e.g*., Laura Weidinger et al., *Taxonomy of Risks Posed by Language Models*, In 2022 ACM Conference on Fairness, Accountability, and Transparency, 214-229 (2022)*;* Dan Hendrycks et al., *Unsolved Problems in ML Safety* (2021); BRIAN CHRISTIAN, THE ALIGNMENT PROBLEM: MACHINE LEARNING AND HUMAN VALUES (2020); Miles Brundage et al., *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, arXiv:1802.07228 (2018); Shahar Avin et al., *Filling Gaps in Trustworthy Development of AI*, Science 374, no. 6573, 1327-1329 (2021); Thomas Arnold & Matthias Scheutz, *The "Big Red Button" Is Too Late: An Alternative Model for The Ethical Evaluation of AI Systems*, Ethics and Information Technology (2018).

[10] *See, e.g.*, James C. Scott, *Seeing Like a State* (1998).

[11] *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,* Utah L. Rev. 1021 (2018); Andrew D. Selbst, *Negligence and AI's Human Users*, BUL Rev. 100 1315 (2020); Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 Wash. L. Rev. 579 (2018).

[12] *See, e.g.*, Andrew Tutt, *An FDA For Algorithms*, Administrative L. REV. Vol. 69, No. 1, 83 (2017) (arguing that a centralized agency is needed for regulating AI); Jack Clark & Gillian K. Hadfield, *Regulatory Markets for AI Safety*, arXiv:2001.00078 (2019); Jonas Schuett, *Defining the Scope of AI Regulations*, Legal Priorities Project Working

behind it (*Law Governs Code*); and how AI can improve the practice of law[13] or implementation of policy[14] (*Code Informs Law*).[15] This Article describes an approach to a third pillar: how law, as theoretical scaffolding and data, can help build more useful and safer AI by design (*Law Informs Code*).

This Article does not assess the legality of AI behavior,[16] or recommend that AI should play a greater (or lesser) role in legal practice;[17] these are critical topics we do not have room to address here. We focus on how AI systems would be more aligned with human preferences and societal values if we couple legal informatics with AI research and deployment.[18] The sociology of finance has advanced the idea that financial economics, conventionally viewed as merely a lens on financial markets, actually shapes markets, i.e., the theory is "an engine, not a camera."[19] Law is an engine, *and* a camera. Legal drafting and interpretation methodologies refined within contract law – an engine of private party alignment – are a lens on how to communicate inherently ambiguous goals of particular humans. Public law – an engine for compliance – is a high-fidelity lens on broader societal values.

Specifying the desirability (i.e., *value*) of an AI system taking a particular *action* in a particular *state* of the world is unwieldy beyond a very limited set of *value-action-state* tuples. In fact, the purpose of machine learning is to train on a subset of tuples[20] and have the resulting agent learn decision policies that generalize to choosing high value actions in unencountered states,[21]

Paper Series No. 9 (2021); Eric Wu et al., *How Medical AI devices Are Evaluated: Limitations and Recommendations From an Analysis of FDA Approvals*, Nature Medicine 27, 4, 582 (2021); Axel Walz & Kay Firth-Butterfield, *Implementing Ethics Into Artificial Intelligence: A Contribution, From A Legal Perspective, To The Development of an AI Governance Regime*, 17 Duke L. & Tech. Rev. (2018).

[13] *See, e.g.*, Henry Prakken, *On How AI & Law Can Help Autonomous Systems Obey the Law: A Position Paper*, AI4J–Artificial Intelligence for Justice 42 (2016) at 44 ("AI & law research has traditionally focused on support tools for humans carrying out legal tasks."); Howard Turtle, *Text Retrieval in the Legal World*, A.I. & L., Vol 3, 5–54 (1995).

[14] *See, e.g.*, Peter Henderson, Ben Chugg, Brandon Anderson & Daniel E. Ho, *Beyond Ads: Sequential Decision-Making Algorithms in Law and Public Policy* (2022); Hannah Bloch-Wehba, *Access to Algorithms*, 88 Fordham L. Rev. 1265 (2019) at 1273 – 1290 (Describes some existing uses of AI by the government.); Emily Berman, *A Government of Laws and Not of Machines*, 98 Bu L. Rev. 1277 (2018).

[15] An emerging, complementary, intersection is *AI as Law*, in which "AI systems are to be thought of as hybrid critical discussion systems, where different hypothetical perspectives are constructed and evaluated until a good answer is found." Bart Verheij, *Artificial Intelligence as Law*, A.I. & L. 28, no. 2, 181-206 (2020) at 191. *AI as Law* comes from the tradition of symbolic systems, *Law Informs AI* is rooted in machine learning, and both are interested in hybrid symbolic-deep learning systems.

[16] *See, e.g.*, Roger Michalski, *How to Sue a Robot*, Utah L. Rev. (2018); Simon Chesterman, *We, The Robots?*. Cambridge University Press (2021).

[17] *See, e.g.*, Frank Pasquale & Glyn Cashwell, *Four Futures of Legal Automation*, 63 U.C.L.A. L. REV. DISC. 26 (2015); Benjamin Alarie, *The Path of the Law: Towards Legal Singularity*, 66 U. TORONTO L.J. 443 (2016); Emily Berman, *A Government of Laws and not of Machines*, 98 Bu L. Rev. 1277 (2018); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, GEO. WASH. L. REV (2019); Aziz Z. Huq, *A Right to a Human Decision*, 106 Va. L. Rev. 611 (2020).

[18] In this Article, we focus on U.S. law, in particular. *See infra* Section V. for a discussion of this limitation.

[19] *See*, Donald MacKenzie, *An Engine, Not a Camera*, MIT Press (2006).

[20] Or input-output pairs, if the focus is purely prediction rather than taking actions. But in this Article we focus on the more general problem of choosing actions, rather than merely prediction; *see, generally*, D. Abel, J. MacGlashan & M.L. Littman, *Reinforcement Learning as a Framework for Ethical Decision Making*, AAAI Workshop: AI, Ethics, and Society (2016). Below, in the context of self-supervised learning, we discuss the increasingly porous distinction between the paradigms of AI prediction (supervised learning) and AI decision-making (reinforcement learning).

[21] Furthermore, a primary purpose of trying to develop future highly advanced AI systems (*see infra* Section II. B.) is to conduct tasks that no human is capable of, *see, generally*, Richard Ngo, *The Alignment Problem From a Deep*

maintaining the same level of performance in novel circumstances sufficiently different from the training data.[22] The reward function ascribing values to an agent's actions during training is inevitably a proxy for the breadth of relevant human preferences,[23] and the training process is unavoidably a sparse exploration of states of all possible futures.[24] In relatively simple, simulated environments, AI systems often exhibit unanticipated "shortcut" behaviors that optimize their inherently limited reward functions,[25] leading the systems to aggressively seek specified rewards at the expense of other (usually less quantifiable) variables of interest.[26] Unintended negative behaviors may result.[27] For instance, when a robot hand was trained to grasp a ball, it instead hovered between the human evaluator's camera and the ball in order to merely give the impression

Learning Perspective (Aug 10, 2022) https://www.alignmentforum.org/posts/KbyRPCAsWv5GtfrbG/the-alignment-problem-from-a-deep-learning-perspective.

[22] Generalization is difficult because machine learning model outputs are effectively interpolations within the model's data manifold, which is defined by the training processes, *see,* François Chollet, *On the Measure of Intelligence* (2019). For discussion of generalization in the context of reinforcement learning, *see, e.g.,* K. Cobbe et al., *Quantifying generalization in reinforcement learning*, ICML (2019); Zhang et al., *A Study on Overfitting in Deep Reinforcement Learning* (2018) at 1 ("the same agents and learning algorithms could have drastically different test performance, even when all of them achieve optimal rewards during training.") Ultimately, whether an AI system's generalizability is adequate depends on how it is deployed and its relation to live decision-making processes; *see, e.g.,* John Nay & Katherine Strandburg, *Generalizability: Machine Learning and Humans-in-the-loop*, *in* BIG DATA LAW (Roland Vogl ed., 2021); Ben Green & Yiling Chen, *The Principles and Limits of Algorithm-in-the-loop Decision Making,* in Proceedings of the ACM on Human-Computer Interaction 3.CSCW 1 (2019); Saleema Amershi, Maya Cakmak, William Bradley Knox & Todd Kulesza, *Power to the people: The role of humans in interactive machine learning*, AI Magazine 35, no. 4 105-120 (2014). For an illustration of the lack of generalizability of deep learning models on a vision task, *see, e.g.,* Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt & Vaishaal Shankar, *Do ImageNet Classifiers Generalize to ImageNet?*, Proceedings of the 36th International Conference on Machine Learning, PMLR 97 5389-5400 (2019).

[23] *See, e.g.,* Langosco et al., *Goal Misgeneralization in Deep Reinforcement Learning* (2022); Amodei et al., *Concrete Problems in AI Safety* (2016).

[24] Without loss of much generality to other AI paradigms, we frame the alignment problem from a decision-making (reinforcement learning) perspective.

[25] *See, e.g.,* W. Bradley Knox et al., *Reward (Mis)design for Autonomous Driving*, arxiv.org (Mar. 11, 2022), https://arxiv.org/abs/2104.13906; Victoria Krakovna et al., *Specification Gaming: The Flip Side of AI Ingenuity*, deepmind.com (Apr. 21, 2020), https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity; Alexander Pan, Kush Bhatia & Jacob Steinhardt, *The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models*, arxiv.org (Feb. 14, 2022), https://arxiv.org/abs/2201.03544 [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*, Artificial Life, 26(2) 274–306 (2020); R. Geirhos et al., *Shortcut learning in deep neural networks,* Nature Machine Intelligence, 2(11) 665–673 (November 2020).

[26] *See*, François Chollet, *Deep Learning with Python, Second Edition* (2021) at 450 ("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."). "Excessive literalism" is another way of describing the issue; *see, e.g.,* "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. 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." quoting Stanford Professor, Stuart Russell (https://www.edge.org/conversation/the-myth-of-ai#26015). *See, e.g.,* Brandon Trabucco et al., *Conservative Objective Models for Effective Offline Model-Based Optimization* (2021).

[27] *See, e.g.,* Jack Clark & Dario Amodei, *Faulty Reward Functions in the Wild.* https://openai.com/blog/faulty-reward-functions/ (2016); Ortega & Maini, *Building safe artificial intelligence: specification, robustness and assurance* (2018) https://medium.com/@deepmindsafetyresearch/building-safe-artificial-intelligence-52f5f75058f1; David Manheim & Scott Garrabrant, *Categorizing Variants of Goodhart's Law* (2018); Rachel L. Thomas & David Uminsky, *Reliance on metrics is a fundamental challenge for AI,* Patterns 3, no. 5 100476 (2022).

that it was grasping the ball.[28] Although, *ex post*, this may seem simple to address with an improved reward function; *ex ante*, careful work from experienced machine learning researchers was not able to design a training process to avoid this.[29] And this is in a tightly controlled environment. Real-world circumstances[30] exacerbate misspecification issues.[31] Take the implementation of simple computational rules applied to data relevant to self-driving cars, for example. When fifty-two programmers were assigned the task of each independently automating simple speed limits, there was "significant deviation in number and type of citations issued [on application of their code to the same real-world data …] this experiment demonstrates that even relatively narrow and straightforward "rules" can be problematically indeterminate in practice."[32] More capable AI will further exacerbate misspecification issues, "achieving higher proxy reward and lower true reward than less capable agents."[33]

Even if we could provide many sources of reward functions, there are countless other relevant goals and world attribute valuations missing.[34] Learning objectives and reward functions based on human feedback and human demonstration can help.[35] But we have many cognitive

[28] *See*, Dario Amodei, Paul Christiano & Alex Ray, *Learning from Human Preferences* (June 13, 2017) https://openai.com/blog/deep-reinforcement-learning-from-human-preferences/; Victoria Krakovna, *Paradigms of AI alignment: components and enablers* (2022).
[29] Another example: an AI agent maximized its provided reward by killing itself at the end of the first level of a simulated environment in order to avoid losing in level two, *see*, William Saunders et al., *Trial without Error: Towards Safe Reinforcement Learning via Human Intervention* (2017). For more examples, *see, e.g.*, Victoria Krakovna, *Specification Gaming Examples in AI* (2022) https://docs.google.com/spreadsheets/d/e/2PACX-1vRPiprOaC3HsCf5Tuum8bRfzYUiKLRqJmbOoC-32JorNdfyTiRRsR7Ea5eWtvsWzuxo8bjOxCG84dAg/pubhtml
[30] *See, e.g.*, A. Rupam Mahmood et al., *Benchmarking Reinforcement Learning Algorithms on Real-World Robots*, In Conference on Robot Learning, 561-591. PMLR (2018).
[31] *See, e.g.*, Steven Kerr, *On the Folly of Rewarding A, While Hoping For B,* Academy of Management Journal 18, no. 4 769-783 (1975); Victoria Krakovna, *Paradigms of AI alignment: components and enablers* (2022); Pan, *Effects of Reward Misspecification*.
[32] Lisa A. Shay, Woodrow Hartzog, John Nelson & Gregory Conti, *Do Robots Dream of Electric Laws? An Experiment in the Law as Algorithm*, Presentation at the We Robot Conference 2013 (Apr. 8–9, 2013), http://www.gregconti.com/publications/201303_AlgoLaw.pdf [cited in Brian Sheppard, *The Reasonableness Machine*, 62 BCL Rev. 2259 (2021)].
[33] Pan, *Effects of Reward Misspecification*, at 1 ("More capable agents often exploit reward misspecifications, achieving higher proxy reward and lower true reward than less capable agents. Moreover, we find instances of phase transitions: capability thresholds at which the agent's behavior qualitatively shifts, leading to a sharp decrease in the true reward. Such phase transitions pose challenges to monitoring the safety of ML systems.")
[34] *See, e.g.*, Roel Dobbe, Thomas Krendl Gilbert & Yonatan Mintz, *Hard Choices in Artificial Intelligence*, Artificial Intelligence, 300 103555 (2021).
[35] *See, e.g.*, Pieter Abbeel, Adam Coates, Morgan Quigley & Andrew Y Ng, *An Application of Reinforcement Learning to Aerobatic Helicopter Flight*, in Advances in Neural Information Processing Systems (2007); Jaedung Choi & Kee-Eung Kim, *Inverse Reinforcement Learning in Partially Observable Environments*, Journal of Machine Learning Research 12, 691–730 (2011); Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel & Stuart J Russell, *Cooperative Inverse Reinforcement Learning*, in Advances in neural information processing systems, 3909–3917 (2016); Dylan Hadfield-Menell et al., *Inverse Reward Design*, in Advances in Neural Information Processing Systems, 6768–6777 (2017); Daniel M. Ziegler, et al., *Fine-tuning Language Models From Human Preferences*, arXiv:1909.08593 (2019); Siddharth Reddy et al., *Learning Human Objectives by Evaluating Hypothetical Behavior*, in Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 8020-8029 (2020); Stiennon et al., *Learning to Summarize with Human Feedback*, In 33 Advances in Neural Information Processing Systems 3008-3021 (2020); Theodore Sumers et al., *Learning Rewards from Linguistic Feedback* (2021); Theodore Sumers et al., *Linguistic Communication As (Inverse) Reward Design*, in ACL Workshop on Learning with Natural Language Supervision (2022); Yuntao Bai et al., *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*, arXiv:2204.05862 (2022) https://arxiv.org/abs/2204.05862.

limitations and biases that corrupt this process,[36] including routinely failing to predict (seemingly innocuous) implications of actions (we believe are) pursuant to our goals;[37] and, unless we can properly augment our feedback and demonstration abilities with trustworthy AI assistants,[38] scaling this process to increasingly advanced AI systems is not promising. The crux of the issue is that, regardless of approach, it is not possible to manually specify or automatically enumerate a discernment of humans' desirability of all actions a system might take.[39] Therefore, after training, the system is deployed with an incomplete map of human preferred territory,[40] and the resulting mismatch between what a human wants and what an AI does is the *human-AI* alignment problem.[41] Considering that multiple humans have preferences on the values of state-action pairs, we must grapple with an even more intractable problem: *society-AI* alignment.[42]

We developed three primary desiderata for a framework to increase human-AI and society-AI alignment.[43] *First*, the framework should have a well-developed theory of alignment. Rather

---

[36] *See, e.g.,* Rohin Shah, Noah Gundotra, Pieter Abbeel & Anca Dragan, *On the Feasibility of Learning, Rather Than Assuming, Human Biases for Reward Inference*, In International Conference on Machine Learning (2019); Geoffrey Irving & Amanda Askell, *AI Safety Needs Social Scientists*, Distill 4.2 e14 (2019). On human cognitive biases more generally, *see, e.g.,* Amos Tversky & Daniel Kahneman, *Judgment under Uncertainty: Heuristics and Biases*, Science 185.4157 1124 (1974).

[37] *See, generally,* Gerd Gigerenzer & Reinhard Selten, eds., *Bounded Rationality: The Adaptive Toolbox*, MIT Press (2002); Sanjit Dhami & Cass R. Sunstein, *Bounded Rationality: Heuristics, Judgment, and Public Policy*, MIT Press (2022).

[38] *See, e.g.,* Paul Christiano, Buck Shlegeris & Dario Amodei, *Supervising strong learners by amplifying weak experts* (2018).

[39] Even if it was possible to specify humans' desirability of all actions a system might take within a reward function that was used for training an AI agent, the resulting behavior of the agent is not only a function of the reward function; it is also a function of the exploration of the state space, *see*, Richard Ngo, *AGI Safety from first principles* (AI Alignment Forum, 2020) at 21-24, https://www.alignmentforum.org/s/mzgtmmTKKn5MuCzFJ. Furthermore, even when the "true" reward function is known, different functions can be equally consistent with the training data; *see, e.g.,* Kareem Amin & Satinder Singh, *Towards Resolving Unidentifiability in Inverse Reinforcement Learning* (2016); Soren Minderman & Stuart Armstrong, *Occam's razor is insufficient to infer the preferences of irrational agents,* in Proceedings of the 32nd International Conference on Neural Information Processing Systems (2018).

[40] Furthermore, there are reinforcement learning systems that learn through interaction with a real-world environment (*see*, J. García & F. Fernández, *A Comprehensive Survey on Safe Reinforcement Learning*, Journal of Machine Learning Research, 16, 1437–1480 (2015); Alex Ray, Joshua Achiam & Dario Amodei, *Benchmarking Safe Exploration in Deep Reinforcement Learning* (2019)) and proposals for general artificial agents that conduct live self-supervised learning as the primary training method (*see, e.g.,* Yan LeCunn, *A Path Towards Autonomous Machine Intelligence* (2022) https://openreview.net/forum?id=BZ5a1r-kVsf). *See,* a reinforcement learning focused proposal in the same vein, *e.g.,* Sekar et al., *Planning to Explore via Self-Supervised World Models,* in Proceedings of the 37th International Conference on Machine Learning, 119, 8583-8592 (2020). If the learning process occurs while the AI agent is taking action in the world, the alignment problem is harder to solve.

[41] Simon Zhuang & Dylan Hadfield-Menell, *Consequences of Misaligned AI*, Advances in Neural Information Processing Systems 33 15763-15773 (2020).

[42] *See infra* Section IV.

[43] This Article is focused on the alignment problem primarily with respect to AI behavior and human values. Another AI alignment problem relates to "corrigibility," *see, generally,* Nate Soares et al., *Corrigibility*, In Workshops at the Twenty- Ninth AAAI Conference on Artificial Intelligence (2015) at 1 (*"*We call an AI system "corrigible" if it cooperates with what its creators regard as a corrective intervention, despite default incentives for rational agents to resist attempts to shut them down or modify their preferences"); *see, also,* Tom Everitt, Daniel Filan, Mayank Daswani & Marcus Hutter, *Self-Modification of Policy and Utility Function in Rational Agents* (2016). On AI alignment focused more directly on existential risks to humanity, *see, e.g.,* Dan Hendrycks & Thomas Woodside, *Open Problems in AI X-Risk* (2022); Victoria Krakovna, *Paradigms of AI alignment: components and enablers* (2022); STUART RUSSELL, HUMAN COMPATIBLE: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF CONTROL (2019). For a discussion of AI alignment specifically related to language agents, *see, e.g.,* Zachary Kenton et al., *Alignment of Language Agents*

than overly simple specifications or vacuous high-level principles with no practical grounding, it should be laden with modular constructs built to handle the ambiguity and novelty inherent in the alignment of (human or artificial) agents. *Second*, in recognizing that alignment is a moving target as AI capabilities advance,[44] the ideal framework should be useful for today's AI systems, but also scale with future potential AI advancements.[45] As AI becomes more capable, the framework should provide benchmarks and solutions calibrated to that higher level of capability – even better if the alignment methods directly benefit from the improvements in general AI capabilities research.[46] *Third*, it should already be rigorously battle-tested – even better if the documentation of the battle-testing has produced reams of data that can be leveraged for machine learning.

Law uniquely fulfills our criteria.[47] Alignment is a problem because we cannot *ex ante* specify rules that fully and provably direct good AI behavior.[48] Similarly, parties to a legal contract cannot foresee every contingency of their relationship,[49] and legislators cannot predict the specific circumstances under which their laws will be applied.[50] That is why much of law is a constellation of appropriately flexible standards.[51] Methodologies for making and interpreting law – where one

(2021). For reviews of the broader domain of the safety of highly advanced AI systems that may be built after further technical advancements, *see, e.g.,* Tom Everitt, Gary Lea & Marcus Hutter, *AGI Safety Literature Review* (2018). For an earlier perspective on AI value alignment (the authors may not describe it as "alignment" research), *see, e.g.,* Daniel S. Weld & Oren Etzioni, *The First Law of Robotics (A Call to Arms)*, AAAI (1994).

[44] As models scale in size, compute, and data inputs, there have been "emergent" breakthroughs in their capabilities; *see, e.g.,* Jason Wei et al., *Emergent Abilities of Large Language Models* (2022); Aarohi Srivastava et al., *Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models* (2022); Jason Wei et al., *Finetuned Language Models Are Zero-Shot Learners*, arxiv.org, (Feb. 8, 2022), https://arxiv.org/abs/2109.01652; Wenlong Huang et al., *Inner Monologue: Embodied Reasoning through Planning with Language Models*, arXiv:2207.05608 (2022) at 8; Deep Ganguli et al., *Predictability and Surprise in Large Generative Models* (2022) at 14 ("large generative models have an unusual combination of high predictability - model capabilities scale in relation to resources expended on training - and high unpredictability - before training a model, it's difficult to anticipate all the inputs it will be subjected to, and what capabilities and outputs it will have. The former drives rapid development of such models while the latter makes it difficult to anticipate the consequences of their development and deployment."); Julien Perolat et al., *Mastering the Game of Stratego with Model-Free Multiagent Reinforcement Learning* (2022) ("Stratego is one of the few iconic board games that Artificial Intelligence (AI) has not yet mastered. [...] Stratego has been a grand challenge for the field of AI for decades [...] DeepNash beats existing state-of-the-art AI methods in Stratego [an extremely complex game] and achieved a yearly (2022) and all-time top-3 rank on the Gravon games platform, competing with human expert players.").

[45] *See,* alignment proposals that focus on scaling to more capable AI, *e.g.,* Leike et al., *Scalable Agent Alignment via Reward Modeling: A Research Direction,* https://arxiv.org/abs/1811.07871 (2018) at 2 ("we describe how reward modeling can be applied recursively: agents trained with reward modeling can assist the user in the evaluation process when training the next agent.").

[46] General technical AI research is often called "capabilities" research by AI alignment and safety researchers in order to distinguish mainstream research focused on improving performance on traditional tasks, e.g., prediction, from AI research focused on alignment, security, or safety outcomes.

[47] Of course, law does not embed all of 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, *see infra* Section IV. 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.

[48] *See, e.g.,* Martin Abadi, Leslie Lamport & Pierre Wolper, *Realizable and unrealizable specifications of reactive systems*, in International Colloquium on Automata, Languages, and Programming (1989); Dan Hendrycks et al., *Aligning AI With Shared Human Values* (2021).

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

[50] *See*, John C. Roberts, *Gridlock and Senate Rules,* 88 Notre Dame L. Rev. 2189 (2012); Brian Sheppard, *The Reasonableness Machine*, 62 BCL Rev. 2259 (2021) [Hereinafter Sheppard *Reasonableness*].

[51] These "standards" are not academic philosophical guidelines. Rather, they are *legal* standards that, at least theoretically, have an "objective" resolution (obtained from a court opinion), *see infra* Section II. A. & Section IV.

set of agents develops specifications for behavior, another set of agents applies the specifications to novel circumstances, and then everyone iterates – have been theoretically refined by legal scholars, attorneys, businesses, legislators, regulators, and courts for centuries. Law is the applied philosophy of multi-agent alignment (*requirement one* of our desired criteria), and legal informatics can be the bridge to instill legal reasoning and knowledge – the language of alignment – within AI systems, helping close the currently widening "gap between social requirements and technical feasibility."[52]

As the state-of-the-art for AI advances, we can set higher bars of demonstrated legal understanding capabilities. If a developer claims their system has advanced capabilities on tasks, then they should demonstrate correspondingly advanced legal comprehension and legal reasoning abilities of the AI,[53] which have practically no ceiling of difficulty when considering the morass of laws and regulations across time, precedent, and jurisdiction (*requirement two*). Benchmarks are the guiding lights of AI research. Legal understanding skills could serve as alignment benchmarks for AI development. No current AI system exhibits the general legal reasoning skills of expert human lawyers, and human experts do not represent the theoretical pinnacle of legal comprehension and reasoning abilities.[54]

The practices of making, interpreting, and enforcing law have been battle tested through millions of legal actions that have been memorialized in digital format[55] that can be leveraged for machine learning (*requirement three*).[56]

Part II expands on the satisfaction of our three requirements to demonstrate why the legal lens is well-suited to increase AI alignment. *Law Informs Code* in two primary ways. *First*, it 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 system to proactively accomplish in the world. This is the topic of Part III. *Second*, by leveraging public law as data, *Law Informs Code* by helping parse what AI systems should generally *not* do, providing an up-to-date distillation of democratically deliberated and determined means toward reducing externalities and promoting societal coordination. We explore this in Part IV.

Significant computing and data resources are required to develop state-of-the-art AI[57] and large companies are key actors pushing the research and deployment boundaries.[58] Regardless of where AI systems are developed (academic labs, private companies, independent groups, etc.),

---

[52] Mark S. Ackerman, *The intellectual challenge of CSCW: the gap between social requirements and technical feasibility*, 15.2-3 Human–Computer Interaction 179 (2000).

[53] Or of other specialized Legal Informatics AI systems that are directly available for guiding the knowledge and actions of the primary AI.

[54] But they can help evaluate advanced AI systems. I cannot do a backflip, but I can evaluate whether you just did one. Furthermore, "One solution is to have humans provide a training signal by demonstrating or judging performance, but this approach fails if the task is too complicated for a human to directly evaluate. We propose Iterated Amplification, an alternative training strategy which progressively builds up a training signal for difficult problems by combining solutions to easier subproblems." Paul Christiano, Buck Shlegeris & Dario Amodei, *Supervising strong learners by amplifying weak experts* (2018) at 1.

[55] *See, e.g.*, Christine Bannan, *Legal Data Access*, *in* LEGAL INFORMATICS (Daniel Martin Katz et al. eds. 2021).

[56] *See, generally*, Daniel Martin Katz & John Nay, *Machine Learning and Law*, *in* LEGAL INFORMATICS (Daniel Martin Katz et al. eds., 2021).

[57] *See, e.g.*, Or Sharir, Barak Peleg & Yoav Shoham, *The Cost of Training NLP Models: A Concise Overview* (2020).

[58] *See e.g.*, this statement by a consortium of AI research companies, Cohere (https://cohere.ai/about), OpenAI (https://openai.com/about/), and AI21 Labs (https://www.ai21.com/about), https://openai.com/blog/best-practices-for-deploying-language-models/.

there is a disconnect between those affected by AI systems and the designers. To increase the alignment of AI systems with the millions of people impacted by them daily, scholars and companies have suggested embedding "ethics" in them.[59] However, it is unclear who decides that "ethics."[60] Our goal, pursued in Part IV, is to reframe the broader multi-agent alignment agenda toward embedding democratically endorsed legal knowledge into AI as the core alignment framework. This complements existing approaches aimed at encoding and embedding ethics, but provides legitimate grounding. Although public law is a reflection of the path-dependent structure of political power within a society and not a perfect aggregation of human values, it is likely the most authoritative encapsulation of the attitudes, norms and values of the governed. We conclude, in Part V, with where further research could be most fruitful, given the possibilities raised.

II. LEGAL INFORMATICS FOR AI ALIGNMENT

A legal (informatics) approach satisfies our three requirements for an alignment framework. Law is the applied philosophy of multi-agent alignment (*II. A*); legal informatics can calibrate AI *task capabilities* and AI *alignment capabilities* as technologies advance (*II. B*); and law produces data and (in the machine learning parlance) model "inductive biases" that can be leveraged to improve AI deployment outcomes through goal communication mechanisms and rich background knowledge on how to act without undue externalities (*II. C*).

A. Legal Theory is Well-Developed, and Applicable to Alignment

Law is a unique discipline – it is both deeply theoretical (even busy practicing attorneys publish esoteric Law Review Articles) and its theories are tested against reality and refined with an unrelenting cadence. Because producing, interpreting, enforcing, and amending law is a never-ending society-wide project,[61] the results are a prime source of information to assist in scalably shaping AI behavior.

1. Law as Information

We are not aiming for AI to have the legitimacy to make law, set legal precedent, or enforce law. In fact, this would undermine our approach. Rather, the most ambitious goal of *Law Informing*

---

[59] *See*, Daniel Greene, Anna Lauren Hoffmann & Luke Stark, *Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning*, in Proceedings of the 52nd Hawaii International Conference on System Sciences (2019); Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol 1, 501–507 (2019) [Hereinafter, Mittelstadt *Principles alone cannot guarantee ethical AI*].

[60] *See, generally*, Mittelstadt *Principles alone cannot guarantee ethical AI*; Frank Pasquale, *New Laws of Robotics: Defending Human Expertise in the Age of AI* (2020).

[61] Law is also capable of reflecting the rights of future generations, *see, e.g.*, Eric Martínez & Christoph Winter, *Protecting Future Generations: A Global Survey of Legal Academics* (2022).

Code is to computationally encode and embed the generalizability of existing legal concepts and standards into validated AI performance. Setting new legal precedent (which, broadly defined, includes proposing and enacting legislation, promulgating agency rules, publishing judicial opinions, enforcing law, and more) should be exclusively reserved for the democratic governmental systems expressing uniquely *human* values.[62] Humans should always be the engine of law-making.[63] The positive implications (for the *Law Informs Code* approach) of that normative stance are that the resulting law then encapsulates human views.

The law is a complex system,[64] with seemingly chaotic underlying behavior, from which aggregated and systematized preferences emerge.[65] Law, leveraged as an expression of *what* humans want,[66] and, critically, *how* they communicate their goals under ambiguity and radical uncertainty,[67] is how *Law Informs Code*. This stands in contrast to more prosaic uses of law, e.g., as an *ex ante* deterrent of bad behavior through the threat of sanction or incapacitation[68] or imposition of institutional legitimacy;[69] or as an *ex post* message of moral indignation.[70] *Law Informs Code* in the tradition of Oliver Holmes and subsequent "predictive" theories of law.[71]

For instance, empirical consequences of violating the law, using enforcement as a source of information,[72] are data points for AI systems. Actually enforcing law on AI systems (or their human creators) is how *Law Governs Code*[73] not how *Law Informs Code*, and is therefore out of

---

[62] *See, e.g.,* Frank Pasquale, *New laws of robotics: defending human expertise in the age of AI* (2020).

[63] *See, e.g.,* Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, GEO. WASH. L. REV (2019).

[64] For a discussion of complex systems science applied to AI Safety, *see, e.g.,* Dan Hendrycks & Thomas Woodside, *Complex Systems for AI Safety* (May 23, 2022) https://www.alignmentforum.org/posts/n767Q8HqbrteaPA25/complex-systems-for-ai-safety-pragmatic-ai-safety-3.

[65] On law as a complex emergent system, *see, e.g.,* Daniel M. Katz & Michael J. Bommarito, *Measuring the complexity of the law: the United States Code*, A.I. & L. 22.4 337-374 (2014); J.B. Ruhl & Daniel M. Katz, *Measuring, Monitoring, and Managing Legal Complexity*, 101 Iowa L. Rev. 191 (2015); Daniel M. Katz et al., *Complex Societies and the Growth of the Law*, Sci Rep 10, 18737 (2020).

[66] Richard H. McAdams, *The Expressive Powers of Law*, Harv. Univ. Press (2017) at 6-7 ("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*]

[67] On the notion of radical uncertainty, *see*, John Kay & Mervyn King, *Radical Uncertainty*, WW Norton & Company (2021); Frank H. Knight, *Risk, Uncertainty, and Profit*, Houghton Mifflin Company (1921).

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

[69] Kenworthey Bilz & Janice Nadler, *Law*, *Psychology & Morality*, in MORAL COGNITION AND DECISION MAKING: THE PSYCHOLOGY OF LEARNING AND MOTIVATION, D. Medin, L. Skitka, C. W. Bauman, & D. Bartels, eds., Vol. 50, 101-131, Academic Press (2009).

[70] Mark A. Lemley & Bryan Casey, *Remedies for Robots*, University of Chicago L. Rev. (2019) at 1347. *See, also, e.g.,* Yuval Feldman, *The Law of Good People: Challenging States' Ability to Regulate Human Behavior* (Cambridge 2018).

[71] Oliver Wendell Holmes, Jr., *The Path of the Law*, in Harvard L. Rev. 10, 457 (1897); Catharine Pierce Wells, *Holmes on Legal Method: The Predictive Theory of Law as an Instance of Scientific Method*, S. Ill. ULJ 18, 329 (1993); Faraz Dadgostari et al. *Modeling Law Search as Prediction*, A.I. & L. 29.1, 3-34 (2021).

[72] McAdams, *The Expressive Powers of Law* at 169-198.

[73] Furthermore, using legal remedies to prevent illegal behavior is difficult with non-human agents, *see*, Mark A. Lemley & Bryan Casey, *Remedies for Robots*, University of Chicago L. Rev. (2019) at 1315, 1316 ("Often, we want to compel defendants to do (or not do) something in order to prevent injury. Injunctions, punitive damages, and even remedies like disgorgement are all aimed—directly or indirectly—at modifying or deterring behavior. But deterring robot misbehavior is going to look very different than deterring humans. [...] Courts, for instance, can rely on the fact that most of us don't want to go to jail, so we tend to avoid conduct that might lead to that result. But robots will be deterred only to the extent that their algorithms are modified to include sanctions as part of the risk-reward calculus.");

the purview of this Article. What good is the law if it is not enforceable – isn't there "no right without a remedy"?[74] From the perspective of an AI system, the law can serve as a rich set of methodologies for interpreting inherently incomplete specifications of collective human expectations.[75] Law provides detailed variegated examples of its application, generalizable precedents with accompanying explanations, and well-trained active lawyers to solicit targeted machine learning model training and fine-tuning feedback to embed an ever-evolving comprehension of human and societal goals.[76] Law, as purely a source to learn goal specification and interpretation[77] methods and (automatically updated and verified) societal knowledge, can play a starring role in aligning AI. We illustrate the applicability of legal theory with three initial examples.

2. Example Theoretical Framing: Complete vs. Incomplete Contracts

The legal lens helps conceptually frame and clarify human-AI alignment. It provides an ontology for alignment.[78] For instance, from this perspective, one way of viewing the alignment of a human with an AI system is the recognition that it is not possible to create a complete contingent "contract" between the AI and the human it serves because the AI training and validation process is not comprehensive of the states of the world that could be encountered after live deployment.[79] This highlights the need for AI to learn modular extra-contractual standards[80] that can generalize across much of the implicit space of potential "contracts."[81]

Ronald Leenes & Federica Lucivero, *Laws on robots, laws by robots, laws in robots: Regulating robot behaviour by design*, Law, Innovation and Technology 6.2, 193 (2014).

[74] Frederick Pollock, *The Continuity of the Common Law*, 11 Harv L Rev 423, 424 (1898).

[75] For more on law as an information source on public attitudes and risks, *see*, Richard H. McAdams, *An Attitudinal Theory of Expressive Law* (2000). For more on law as a coordinating mechanism, *see*, Richard H. McAdams, *A Focal Point Theory of Expressive Law* (2000).

[76] *See infra* Section III. B.

[77] *See, e.g.,* Owen M. Fiss, *Objectivity and Interpretation*, 34 STAN. L. REV. 739 (1982).

[78] *See, e.g.,* Arbital, *Ontology Identification Problem* https://arbital.com/p/ontology_identification/. *See*, a legal ontology, *e.g.*, P Casanovas et al., *Semantic Web for the Legal Domain: The Next Step*, Semantic Web 7(3):213–227 (2016).

[79] For the contract-AI alignment analogy, *see*, Dylan Hadfield-Menell & Gillian K. Hadfield, *Incomplete Contracting and AI Alignment*, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019) at 422, 471 (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." In contrast to Hadfield-Menell & Hadfield (2019), who conclude that the primary need is to build "AI that can replicate human cognitive processes", 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.)

[80] *See infra* Section II. A. 3. & Section III.

[81] An Inverse Reinforcement Learning artificial "agent might not ever learn what is the best (or the morally or ethically appropriate) action in some regions of the state space. Without additional capabilities, it would be incapable of reasoning about what ought to be done in these regions – this is exactly the reason why we have norms in the first place: to not have to experience all state/actions precisely because some of them are considered forbidden and should not be experienced." Thomas Arnold et al., *Value Alignment or Misalignment - What Will Keep Systems Accountable?* AAAI Workshops (2017) at 5.

3. Example Theoretical Framing: Rules vs. Standards

Another example of the legal lens helping frame AI alignment is legal theory concerning the distinction between rules and standards.[82] Rules (e.g., "do not drive more than 60 miles per hour") 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 are a function of any behavioral impact the rules might have had).[83] Complex social systems have emergent behavior that can make formal rules brittle.[84] Standards (e.g., "drive reasonably") allow parties to contracts, judges, regulators, and citizens to develop shared understandings and adapt them to novel situations, i.e., to generalize expectations regarding actions taken to unspecified states of the world. If rules are not written with enough potential states of the world in mind, they can lead to unanticipated undesirable outcomes[85] (e.g., a driver following the rule above is too slow to bring their passenger to the hospital in time to save their life), but to enumerate all the potentially relevant state-action pairs is excessively costly outside of the simplest environments.[86]

A standard has the capacity to be more generalizable to novel situations than specific rules.[87] The AI analogy for legal standards are continuous, approximate methods that rely on significant amounts of data for learning dense representations on which we can apply geometric operations in the model's latent space. They are flexible.[88] The AI analogy for rules is human-crafted "if-then" statements that are brittle, yet require no empirical data for machine learning.[89]

---

[82] *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); 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. (2014); Michael Coenen, *Rules Against Rulification*, 124 YALE L.J. (2014); Anthony J. Casey & Anthony Niblett, *Death of Rules and Standards,* 92 IND. L.J. (2017); Sheppard *Reasonableness*.

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

[84] *See, e.g.,* Dylan Hadfield-Menell, McKane Andrus & Gillian Hadfield, *Legible Normativity for AI alignment: The Value of Silly Rules*, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 115-121 (2019).

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

[86] *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.*

[87] *See, e.g.,* the SEC explains 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." *Commission Interpretation Regarding Standard of Conduct for Investment Advisers* at 5. *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,* in The Journal of Corporation Law (2017).

[88] Patterns of legal language in contracts exhibit elasticity, *see,* Klaudia Galka & Megan Ma, *Measuring Contract Elasticity: Computing Reinsurance* (Stanford Law School, 2022); Grace Q. Zhang, *Elastic Language: How and Why We Stretch Our Words* (2015).

[89] Harry Surden, *The Variable Determinacy Thesis*, 12 COLUMB. SCI. & TECH. L. REV. 1 (2011). [Hereinafter, Surden, *Variable Determinacy*].

In practice, most legal provisions land somewhere on a spectrum between pure rule and pure standard,[90] and legal theory can help us estimate the right location and combination[91] when building AI systems. Furthermore, there are other theorized dimensions to legal provision implementation related to the "rule-ness" versus "standard-ness" axis that could further elucidate AI design, e.g., "determinacy,"[92] "privately adaptable" ("rules that allocate initial entitlements but do not specify end-states"[93]), 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""[94]).

4. Example Theoretical Framing: Private vs. Public Law

The alignment problem is most often described (usually implicitly) in the AI research literature with respect to the alignment of one AI system with one human, or a small subset of humans.[95] It is more challenging to expand the scope of the AI's analysis beyond a small set of humans and ascribe *societal value* to action-state pairs.[96] The legal framing highlights differences between addressing *human-AI* alignment and *society-AI* alignment.[97] The latter requires us to move beyond private contracts and into the realm of public law to explicitly address inter-agent relationships and policies designed to ameliorate externalities and solve massively multi-agent coordination and cooperation dilemmas.[98]

---

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

[91] *See*, Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 7 Columbia L. Rev. 119, 1851 (2019) at 1859 ("Decision criteria may also combine rule-like and standard-like aspects according to various schemes. For example, DWI laws in many states combine a rule-like blood alcohol threshold, above which a finding of intoxication is required, with a standard-like evaluation of intoxication at lower levels. Some speed limit laws use a somewhat different scheme: Above a rule-like speed limit, there is a presumption of unsafe driving, but adjudicators may make standard-like exceptions for a narrow range of emergency circumstances.").

[92] Surden, *Variable Determinacy.*

[93] Cass R. Sunstein, *Problems with Rules*, 83 CALIF. L. REV. 953 (1995) at 959.

[94] Gideon Parchomovsky & Alex Stein, *Catalogs*, 115 COLUM. L. REV. 165 (2015) at 165.

[95] *See, e.g.,* Amanda Askell et al., *A General Language Assistant as a Laboratory for Alignment* (2022) at 44 ("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.") [Hereinafter Askell Laboratory for Alignment]; Stiennon et al., *Learning to Summarize with Human Feedback,* In 33 Advances in Neural Information Processing Systems 3008-3021 (2020). For a high-level overview of AI alignment research, *see,* Jan H. Kirchner et al., *Researching Alignment Research: Unsupervised Analysis* (2022).

[96] *See, e.g.,* Jiaying Shen, Raphen Becker & Victor Lesser, *Agent Interaction in Distributed POMDPs and its Implications on Complexity*, in Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, 529-536 (2006).

[97] As one moves from private to public law, that broadens "human-human alignment" challenges. Sophisticated general human reasoning capabilities, e.g., a general counsel's office with teams of specialists on call with decades of experience in niche areas of law, are deployed in attempts to navigate the challenges associated with understanding and complying with public law.

[98] *See, e.g.,* Elinor Ostrom, *Understanding Institutional Diversity*, Princeton University Press (2009); Phillip Christoffersen, Andreas A. Haupt & Dylan Hadfield-Menell, *Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*, arXiv:2208.10469 (2022).

B. Legal Informatics Can Scale with AI Capabilities

Within the *Law Informs Code* framework, we refer to the ability of AI to perform narrow tasks for humans as "AI-contract" capability. This level of capability is already widely deployed, e.g., through powering billions of automated online advertisement placements and social media content choices every day.[99] Large neural-network-based models pre-trained with self-supervision[100] on significant portions of the internet that require little to no traditional supervised learning to perform well on some new tasks ("Foundation Models"[101]) are beginning to display what we call "AI-standards" capabilities, which could be used to help align AI with humans through legal informatics as described herein.[102] Standards are more abstract and nuanced than

---

[99] *See, e.g.,* recent earnings calls (July 2022) where Meta (Facebook) CEO Mark Zuckerberg discusses how AI-driven recommendations from accounts a user does not follow represents 15% of a user's content in Facebook and more in Instagram, and will more than double by the end of 2023. *See, e.g.,* Nathan Baschez, *Instagram's Existential Bet* (2022) https://every.to/divinations/instagram-s-existential-bet ("In that same earning call Zuckerberg talked about how there's a "major shift" towards discovery algorithms, because it "unlocks a large amount of interesting and useful videos and posts you might have otherwise missed." But he doesn't just see this as unique to Reels / TikTok. He wants to turn this into a major drive for all types of content across the Facebook family of apps, saying "I think about the AI that we're building not just as a recommendation system for short-form video, but as a Discovery Engine that can show you all of the most interesting content that people have shared across our systems."). *See, generally*, Jonathan Stray, et al., *Building Human Values into Recommender Systems: An Interdisciplinary Synthesis* (2022) https://arxiv.org/abs/2207.10192; Jonathan Stray et al., *What are you optimizing for? Aligning recommender systems with human values*, arXiv 2107.10939 (2021).

[100] "Self-supervised" training procedures predict data items that are systematically held-out, e.g., removing a word and predicting the word that was removed (and then iterating this billions of times), or predicting whether an entire sentence is next to another sentence. No explicit labeling of the data is required; therefore, self-supervised training allows a model to be trained across significantly more data than traditional supervised learning. Scalable self-supervised pruning of the training data can reduce the costs of training, *see, e.g.,* Ben Sorscher et al., *Beyond neural scaling laws: beating power law scaling via data pruning* (2022).

[101] *See, e.g.,* Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) https://arxiv.org/pdf/2108.07258.pdf. Foundation Models leverage applications of the Transformer architecture (*see, e.g.,* Ashish Vaswani et al., *Attention Is All You Need*, *in* PROCEEDINGS OF THE 31ST CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (2017)), which is a model used to encode an input sequence (e.g., words in a particular order) into context-aware representation and then decode that into a novel generation of an ordered sequence (e.g., a new set of words in a particular order) as an output. Sequences of words were the first application area of this model architecture with major success, as these models can capture complicated dependencies and interactions within natural language (LEWIS TUNSTALL ET AL., NATURAL LANGUAGE PROCESSING WITH TRANSFORMERS, xii (2022)). The Transformer has since been applied beyond natural language, also to graphs (*see, e.g.,* Jinwoo Kim et al., *Pure Transformers are Powerful Graph Learners* (2022)) and decision-making (*see, e.g.,* Micah Carroll et al., *Towards Flexible Inference in Sequential Decision Problems via Bidirectional Transformers*, (2022) https://arxiv.org/abs/2204.13326). Even within the natural language data structure, Transformer models demonstrate non-language skills, such as mathematical reasoning skills, *see, e.g.,* Aitor Lewkowycz et al., *Solving Quantitative Reasoning Problems with Language Models* (2022) https://arxiv.org/abs/2206.14858; Tuan Dinh et al., *LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks* (2022) https://arxiv.org/abs/2206.06565; and the ability to simulate social systems, *see, e.g.,* Joon Sung Park et al., *Social Simulacra: Creating Populated Prototypes for Social Computing Systems*, in 35th Annual ACM Symposium on User Interface Software and Technology (2022).

[102] The AI systems most capable of performing well on diverse sets of tasks – the most generalizable – are large neural network based models trained on large, diverse data sets through self-supervision (Foundation Models); *see, e.g.,* Tom Brown et al., *Language models are few-shot learners*, Advances in Neural Information Processing Systems (2020); Alexander Kolesnikov et al., *Big transfer (bit): General visual representation learning,* ECCV (2020); Thoppilan et al., *LaMDA: Language Models for Dialog Applications* (2022); Tran et al., *Plex: Towards Reliability Using*

rules, and require more generalizable capabilities and world knowledge to implement. The next level – somewhere between current Foundation Model capabilities and a potential "Artificial General Intelligence"[103] (AGI) capability level, *paired with* the additional development of methods and datasets specific to legal understanding[104] – may unlock what we can call an "AI-public" capability, where an AI system and its legal-informatics-inspired architecture and training process is able to understand flexible standards, interpretation guidelines, legal reasoning, and generalizable precedents (which, effectively, synthesize citizens' value preferences over potential actions taken in many states of the world).[105] We are not there yet.

1. AI Capabilities

Unlike traditional supervised learning,[106] self-supervision, reinforcement learning, and imitation learning can lead to "emergent" capabilities.[107] Advancements in deep reinforcement learning are producing more generalizable decision-making agents,[108] which lead to more pressing alignment issues due to the possibility of more autonomous deployments. For both pure prediction

---

Pretrained Large Model Extensions (2022) https://ai.googleblog.com/2022/07/towards-reliability-in-deep-learning.html.

[103] *See, e.g.,* a definition from OpenAI, "highly autonomous systems that outperform humans at most economically valuable work" https://openai.com/charter/ (accessed Aug 23, 2022). *See, generally,* Tom Everitt, Gary Lea & Marcus Hutter, *AGI Safety Literature Review* (2018); Richard Ngo, *The Alignment Problem From a Deep Learning Perspective* (Aug 10, 2022) https://www.alignmentforum.org/posts/KbyRPCAsWv5GtfrbG/the-alignment-problem-from-a-deep-learning-perspective.

[104] *See, generally, infra* Section II. B. 2.

[105] There are no clear boundaries between the AI-public, AI-standards, and AI-contract levels. We are simply using them as high-level rhetorical devices.

[106] With supervised learning, the best one can do is to predict, with perfect calibration, the same type of output variable with the same type of input variables on new test data not used for training the model. A model that can generalize to novel tasks is out of scope of the traditional supervised learning framework, which is concerned only with mapping *ex ante*-specified input types to *ex ante*-specified output types, albeit on unseen input data points. A complication to this characterization of the distinction in learning paradigms, though, is when supervised learning is trained on state-action-reward sequences and deployed to generate actions in new states by conditioning on high rewards, *see, e.g.,* Dylan R Ashley et al., *Learning Relative Return Policies With Upside-Down Reinforcement Learning*, arXiv:2202.12742 (2022); Kai Arulkumaran et al., *All You Need Is Supervised Learning: From Imitation Learning to Meta-RL With Upside Down RL*, arXiv:2202.11960 (2022).

[107] *See, e.g.,* Ganguli et al., *Predictability and Surprise in Large Generative Models* (2022) at 2 https://arxiv.org/abs/2202.07785 ("Our basic thesis is that large generative models have an unusual combination of high predictability — model loss improves in relation to resources expended on training, and tends to correlate loosely with improved performance on many tasks — and high unpredictability — specific model capabilities, inputs, and outputs can't be predicted ahead of time. The former drives rapid development of such models while the latter makes it difficult to anticipate the consequences of their development and deployment."); Wenlong Huang et al., *Inner Monologue: Embodied Reasoning through Planning with Language Models*, arXiv:2207.05608 (2022) at 8 https://arxiv.org/abs/2207.05608 ("Although LLMs can generate fluent continuation from the prompted examples, we surprisingly find that, when informed with environment feedback, Inner Monologue demonstrates many impressive reasoning and replanning behaviors beyond the examples given in the prompt. Using a pre-trained LLM as the backbone, the method also inherits many of the appealing properties from its versatility and general-purpose language understanding.").

[108] *See e.g.,* Bhoopchand et al., *Learning Robust Real-Time Cultural Transmission without Human Data* (2022) at 2 ("Our artificial agent is parameterised by a neural network and we use deep reinforcement learning (RL) to train the weights. The resulting neural network (with fixed weights) is capable of zero- shot, high-recall cultural transmission within a "test" episode on a wide range of previously unseen tasks.")

tasks (e.g., typical supervised learning inference) and decision tasks (e.g., deployment of agents trained through reinforcement or imitation learning), the process of first conducting self-supervised training on a large scale, has proven effective for performing well on downstream tasks, oftentimes with little fine-tuning on those tasks – for "few-shot reasoning."[109] With latent parameter spaces and "concepts" learned through pre-training, models can even exhibit robust capabilities on novel tasks without any fine-tuning on that particular task – through techniques such as in-context learning,[110] manually engineered prompting,[111] and automated algorithmic prompting[112] – for "zero-shot reasoning."[113] Foundation Models currently excel in natural language processing, but they are also being successfully applied beyond textual data.[114] Large

[109] *See, e.g.*, Yaqing Wang et al., *Generalizing from a few examples: A survey on few-shot learning*, ACM computing surveys (csur) 53.3 1-34 (2020); Zhao Mandi, Pieter Abbeel & Stephen James, *On the Effectiveness of Fine-tuning Versus Meta-reinforcement Learning* (2022); Victor Sanh et al., *Multitask Prompted Training Enables Zero-Shot Task Generalization* (2022); Jesse Michael Han et al., *Proof Artifact Co-training for Theorem Proving with Language Models*, in The First Mathematical Reasoning in General Artificial Intelligence Workshop, ICLR (2021) https://mathai-iclr.github.io/papers/papers/MATHAI_23_paper.pdf (Self-supervised training helps an AI system with proving mathematical theorems); Dinglan Peng et al., *How Could Neural Networks Understand Programs?* (2021) https://arxiv.org/abs/2105.04297 (Self-supervised training helps an AI system with understanding computer programming).

[110] For potential explanations of how this surprising behavior is possible, *see, e.g.*, Sang Michael Xie et al., *An Explanation of In-context Learning as Implicit Bayesian Inference* (2021); Sewon Min et al., *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* https://arxiv.org/abs/2202.12837 (2022); Sang Michael Xie & Sewon Min, *How does in-context learning work? A framework for understanding the differences from traditional supervised learning*, Stanford AI Lab Blog (August 1, 2022); David Dohan et al., *Language Model Cascades*, arXiv:2207.10342 (2022) https://arxiv.org/abs/2207.10342. In-context learning, or "prompting" (see next footnote) is an example of surprising behavior from powerful AI models, *see*, Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) at 116 https://arxiv.org/pdf/2108.07258.pdf ("Similarly, small rewordings of prompts can have large impacts on task performance. Since the space of prompts is intractable to enumerate, it is challenging to definitely assert that any task is outside the reach of current prompt-based foundation models — this is a major challenge for reasoning about possible catastrophic risks from foundation models.")

[111] *See, e.g.*, Jason Wei et al., *Chain of thought prompting elicits reasoning in large language models*, arXiv:2201.11903 (2022); Adi Robertson, *PROFESSIONAL AI WHISPERERS HAVE LAUNCHED A MARKETPLACE FOR DALL-E PROMPTS: AI art isn't just an experiment — it's a side hustle* (Sep 2, 2022) https://www.theverge.com/2022/9/2/23326868/dalle-midjourney-ai-promptbase-prompt-market-sales-artist-interview (a marketplace for "prompt engineers").

[112] *See, e.g.*, Antonia Creswell & Murray Shanahan, *Faithful Reasoning Using Large Language Models* (2022) https://arxiv.org/abs/2208.14271 at 15 ("Our approach exemplifies a trend towards algorithmic prompting, a form of automated prompt engineering in which querying a language model becomes a computational primitive. The responses of the language model can be manipulated to construct new prompts that are then used to make further queries. Model queries and prompt construction are composed into algorithms with the usual computational constructs: sequence, choice, and iteration. Algorithmic prompting can be used to elicit more sophisticated and nuanced behaviour from a language model than would otherwise be possible. For example, as our work shows, this approach can be used to develop models capable of faithful reasoning, without compromising performance.").

[113] *See, e.g.*, Kojima et al., *Large Language Models are Zero-Shot Reasoners* (2022) at 1 ("we show that LLMs are decent zero-shot reasoners by simply adding "Let's think step by step" before each answer [...] The versatility of this single prompt across very diverse reasoning tasks hints at untapped and understudied fundamental zero-shot capabilities of LLMs, suggesting high-level, multi-task broad cognitive capabilities may be extracted by simple prompting."); K. Bostrom, Z. Sprague, S. Chaudhuri & G. Durrett, *Natural Language Deduction Through Search Over Statement Compositions*, arXiv:2201.06028 (2022); E. Zelikman, Y. Wu & N. D. Goodman, *Star: Bootstrapping Reasoning with Reasoning*, arXiv:2203.14465 (2022) https://arxiv.org/abs/2203.14465; Jang et al., *BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning* (2021).

[114] *See, e.g.*, Kevin Lu et al., *Pretrained Transformers as Universal Computation Engines* (2021) (They "find language-pretrained transformers can obtain strong performance on a variety of non-language tasks."); Lili Chen et al., *Decision Transformer: Reinforcement Learning via Sequence Modeling* (2021) (They abstract reinforcement

models are being released open source, further democratizing their use and distributing their impact.[115]

Increased capabilities and generalizability of AI models could allow us to better align them with humans and with society more broadly, if the capabilities are harnessed in the right way.[116] If AGI is attained, according to the more pessimistic AI alignment theorists, "*Seemingly "simple" proposals [for ensuring super-intelligent systems realize a positive outcome for humans] are likely to have unexpected undesirable consequences, overlooked as possibilities because our implicit background preferences operate invisibly to constrain which solutions we generate for consideration.[…] There is little prospect of an outcome that realizes even the value of being interesting, unless the first superintelligences undergo detailed inheritance from human values*."[117] Inheriting an understanding of legal reasoning, legal interpretation methods, legal standards, and public law could provide future advanced AI systems a more comprehensive view of what humans want. Although we may be decades from AGI capabilities, if they are ever obtained, addressing safety and alignment risks earlier in the deployment lifecycle of incredibly powerful technologies, such as AI, likely leads to better outcomes.[118] Therefore, we propose a legal understanding verification process that *scales with* AI task capabilities, calibrated to the difficulty level of the

---

learning as a sequence modeling problem like language modeling and then leverage a Transformer architecture to condition on the desired reward, past states, and actions, to generate future actions.); Kuang-Huei Lee et al., *Multi-Game Decision Transformers* (2022); Mengdi Xu et al., *Prompting Decision Transformer for Few-Shot Policy Generalization*, in Proceedings of the 39th International Conference on Machine Learning, PMLR 162 24631-24645 (2022); Micah Carroll et al., *Towards Flexible Inference in Sequential Decision Problems via Bidirectional Transformers* (2022) (They develop and test "a framework for flexibly defining and training models which: 1) are naturally able to represent any inference task and support multi-task training in sequential decision problems, 2) match or surpass the performance of specialized models after multi-task pre-training, and almost always surpasses them after fine-tuning."); Victor Sanh et al., *Multitask Prompted Training Enables Zero-Shot Task Generalization* (2022) (They develop and test a system for converting any natural language tasks into a human-readable prompt form.).

[115] *See, e.g.*, Melissa Heikkilä, *BLOOM: Inside the Radical New Project to Democratize AI*, In MIT Technology Review (2022) (A large language model matching in scale some of the best performing large models from OpenAI, Google, and others was trained and released open source in July 2022 by over 1,000 volunteers.); *CodeGen* https://github.com/salesforce/CodeGen ("CodeGen is an open-source model for program synthesis [that is] competitive with OpenAI Codex."); *Stable Diffusion* https://github.com/CompVis/stable-diffusion ("Stable Diffusion is a latent text-to-image diffusion model. […] Similar to Google's Imagen, this model uses a frozen CLIP ViT-L/14 text encoder to condition the model on text prompts […] the model is relatively lightweight and runs on a GPU with at least 10GB VRAM."); *Hugging Face T5* https://huggingface.co/docs/transformers/model_doc/t5 (Originally at Google, "T5 is an encoder-decoder model pre-trained on a multi-task mixture of unsupervised and supervised tasks and for which each task is converted into a text-to-text format."); Thomas I. Liao, *Foundation Model Tracker* (2022) https://foundationmodeltracker.notion.site/foundationmodeltracker/Model-Tracker-v0-9-794ba77f74ec469186efdbdb87e9b8e6.

[116] *See*, Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research* (2022) at 8 ("Improving an agent's world model makes them more generally capable, but this also can make them less likely to spawn unintended consequences. Optimizers operating over longer time horizons will be able to accomplish more difficult goals, but this could also make models act more prudently and avoid taking irreversible actions.").

[117] Eliezer Yudkowsky, *Complex Value Systems are Required to Realize Valuable Futures*, In Artificial General Intelligence: 4th International Conference, AGI 2011, Proceedings edited by Jürgen Schmidhuber, Kristinn R. Thórisson & Moshe Looks, 388–393, Vol. 6830, (2011) at 14 https://intelligence.org/files/ComplexValues.pdf.

[118] *See, e.g.,* Dan Hendrycks & Mantas Mazeika, *X-Risk Analysis for AI Research* (2022) at 6 ("Many early Internet protocols were not designed with safety and security in mind. Since safety and security features were not built in early, the Internet remains far less secure than it could have been, and we continue to pay large continuing costs as a consequence."); *See, generally,* Nancy G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety*, MIT Press (2016).

legal reasoning and interpretation tasks from "AI-contract," to "AI-standards," to "AI-public" capability.

2. Legal Understanding Demonstrations

In most existing applications, before AI models are deployed, their performance on the task at hand is validated on data not employed in their design or training in order to demonstrate generalizability (task performance in circumstances sufficiently different than the training data that is commensurate with task performance on the training data). This out-of-sample performance evaluation is as a demonstration of a capability related to a specific human's (or organization's) preferences. In our framework, this is a demonstration of the AI's "understanding" of the terms of an (implied) "contract" between the AI and the human(s) it is conducting tasks for.

If a system is only at an "AI-contract" capability level, and deployed only on a narrow task, we have no expectation of it being able to autonomously track and comply with public law. This is the status quo for most deployments. It does not mean that the systems will necessarily violate the law; rather, it's an admission that the AI is not advanced enough to understand the legal context in any meaningful sense.[119]

AI deployment is now advancing faster than AI legal understanding because there is vastly more effort on the former.[120] To address the gap,[121] before AI models are deployed in increasingly agentic capacities, e.g., fully autonomous vehicles on major roads,[122] the deploying party should

[119] *See, e.g.,* John Nay & James Daily, *Aligning Artificial Intelligence with Humans through Public Policy* (2022). [Hereinafter Nay, *Aligning AI through Public Policy*].

[120] *See, e.g.,* Dan Hendrycks & Thomas W, *Introduction to Pragmatic AI Safety* (May 9, 2022) https://www.alignmentforum.org/posts/bffA9WC9nEJhtagQi/introduction-to-pragmatic-ai-safety-pragmatic-ai-safety-1 ("Machine learning has been outpacing safety. […] today, we have models capable of answering bar exam questions, writing code, and explaining jokes. Meanwhile, existing approaches to AI safety have not seen similar strides. Many older approaches are still pre-paradigmatic, uncertain about what concrete research directions should be pursued and still aiming to get their bearings. Centered on math and theory, this research focuses on studying strictly futuristic risks that result from potential systems. Unfortunately, not much progress has been made"). This general technological governance issue is often framed as the "pacing problem," *see, e.g.,* Gary E. Marchant, *Governance of Emerging Technologies as a Wicked Problem*, Vand. L. Rev. 73 1861 (2020); Adam Thierer, *The Pacing Problem and the Future of Technology Regulation* (2018).

[121] For other recommendations of potential policy responses , *see, e.g.,* Ryan Calo, *Artificial intelligence policy: a primer and roadmap*, UCDL Rev. Vol. 51 (2017); Jessica Fjeld et al., *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-Based Approaches to Principles for AI* (2020) at 34. For a dashboard of international AI policies, *see* the OECD AI's live repository of over 260 AI policies. For specific recent federal U.S. government proposals, *see, e.g.,* Aiming for truth, fairness, and equity in your company's use of AI, Federal Trade Commission, https://www.reginfo.gov/public/do/eAgendaViewRule?pubId=202110&RIN=3084-AB69; *and*, EEOC Launches Initiative on Artificial Intelligence and Algorithmic Fairness, US Equal Employment Opportunity Commission; *and*, Agencies Seek Wide Range of Views on Financial Institutions' Use of Artificial Intelligence; *and*, Federal Register, Artificial Intelligence Risk Management Framework.

[122] National Conference of State Legislatures, *Autonomous Vehicles | Self-Driving Vehicles Enacted Legislation*, ncsl.org (Feb. 18, 2020), https://www.ncsl.org/research/transportation/autonomous-vehicles-self-driving-vehicles-enacted-legislation.aspx.

demonstrate[123] the system's understanding of human goals, policies, and legal standards.[124] A validation procedure could illustrate the AI's "understanding."[125]

In addition to demonstrating its ability to uphold simulated private contractual obligations through acceptable out-of-sample task performance, sufficiently capable AI systems should demonstrate an ability to perform consistent with extra-contractual standards, such as a fully automated investment advisory system exhibiting simulated behavior in line with fiduciary duties to its human principal. Further, sufficiently agentic AI systems[126] should demonstrate comprehension of the public law that will be relevant to its behavior if it is deployed.[127] This is a very difficult threshold to pass.[128]

As the state-of-the-art for AI advances, there should be a higher bar of demonstrated legal understanding abilities.[129] If an AI developer claims their system has advanced capabilities on tasks that it would like the AI to complete outside of its training environment, the developer should show correspondingly advanced legal knowledge and legal reasoning abilities of the system.

One of the technical impediments to machine learning alignment, and deep learning in particular,[130] is the difficulty of understanding the model and the causes of specific behaviors once

---

[123] Demonstrate to governments, ideally. *See, e.g.*, Jess Whittlestone & Jack Clark, *Why and How Governments Should Monitor AI Development*, arXiv:2108.12427 (2021).
[124] *See infra* Section III.
[125] We do not want to understate the extreme difficulty of developing robust verification processes. More deterministic and much simpler software cannot be fully trusted, *see, e.g.*, Ken Thompson, *Reflections on Trusting Trust*, 27 CACM 761 (1984) https://archive.ph/gi1W8 ("You can't trust code that you did not totally create yourself. [...] No amount of source-level verification or scrutiny will protect you from using untrusted code.").
[126] For a description of advanced AI capabilities and agentic planning, *see, e.g.*, Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?* (2022) at 4-7.
[127] *See infra* Section IV; Nay, *Aligning AI through Public Policy*.
[128] *See, e.g.*, Daniel J. Gervais, *Towards an Effective Transnational Regulation of AI*, AI & Society 1-20 (2021) at 6 ("Take just this well-known example: Carlsbad Technology, Inc. v. HIF Bio, Inc, 556 U.S. 635 (2008). Looking at the statute involved in that case (28 U. S. C. §1441(c)), would lead to an entirely incorrect understanding of "the law" because the Supreme Court's interpretation of the statute—the exact opposite of what the text of the statute actually says—is what courts are bound to follow under stare decisis."). *See*, avenues of AI research focused on resolving conflicting norms, *e.g.*, Daniel Kasenberg & Matthias Scheutz, *Inverse Norm Conflict Resolution*, in Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 178-183 (2018).
[129] Scholars have suggested we use AI "Guardians" to monitor operational AI systems once they are deployed, *see, e.g.*, Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 Vanderbilt Journal of Entertainment and Technology Law 133 (2016) at 139, "From here on, AI should be divided into two categories. The first category would consist of operational AI programs - the computerized "brains" that guide smart instruments. The second category would be composed of oversight AI programs that review the first category's decision making and keep the decisions in line with the law. These oversight programs, which this Article calls "AI Guardians," would include AI programs to interrogate, discover, supervise, audit, and guarantee the compliance of operational AI programs." Our proposal is focused on AI systems demonstrating their own legal understanding, but our discussion of seeking parity between AI capabilities and AI legal understanding is related to what would be required to technically implement the Etzionis' proposal, at 146, "These AI Guardians will need to become smarter just as operational Al programs are improving."
[130] Tilman Räukur, Anson Ho, Stephen Casper & Dylan Hadfield-Menell, *Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks*, arXiv:2207.13243 (2022) https://arxiv.org/abs/2207.13243.

it is trained.[131] This generally gets worse as models scale in size and complexity.[132] Understanding the inner workings of the models is helpful for editing beliefs and facts,[133] and validating its performance, safety, reliability, and legal abilities[134] across potential scenarios,[135] and for enabling humans to potentially take real-time corrective actions.[136] Training deep learning models on legal data, where the learned intermediate representations can in some cases correspond to legal concepts, opens the possibility for mechanistic (alignment) interpretability,[137] methods reverse engineering AI models for better understanding their (misalignment) tendencies.[138] If legal concepts are the ontology for alignment, viewing their use inside trained AI systems helps unpack a mechanistic explanation for how and why a model is aligned.

We should supplement a mechanistic explanation with a behavioral perspective on understanding AI. Simulations exploring the actions of machine-learning-based decision-making

---

[131] *See, generally,* M. Danilevsky et al., *A Survey of the State of Explainable AI for Natural Language Processing* (2020). For some of the latest methods for interpretability of large generative machine learning models, *see, e.g.*, Ganguli et al., *Predictability and Surprise in Large Generative Models* https://dl.acm.org/doi/abs/10.1145/3531146.3533229 (2022). For a counter-argument for the need for explainability in the context of autonomous systems acting legally, *see, e.g.,* Henry Prakken, *On How AI & Law Can Help Autonomous Systems Obey the Law: A Position Paper*, AI4J–Artificial Intelligence for Justice 42, 42-46 (2016) at 44 ("the legal tasks supported by traditional AI & law tools require explanation and justification of decisions. With autonomous systems there is no need for this; all that counts is that legally acceptable behaviour is generated. Of course, when an autonomous system does something legally wrong, its be- haviour might have to be explained in a court case. However, this does not require that the system itself can do that; it may suffice to have a log file recording the system's internal actions."); *see, also, e.g.*, Michael Kearns & Aaron Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, 2019) at 170-175 (Defining "interpretability" depends on the audience and the model type and task.).

[132] Info. Law Inst. at N.Y. Univ. Sch. of Law with Foster Provost, Krishna Gummadi, Anupam Datta, Enrico Bertini, Alexandra Chouldechova, Zachary Lipton & John Nay, *Modes of Explanation in Machine Learning: What Is Possible and What Are the Tradeoffs?*, in Algorithms and Explanations (Apr. 27, 2017), https://youtu.be/U0NsxZQTktk.

[133] Kevin Meng, David Bau, Alex Andonian & Yonatan Belinkov, *Locating and Editing Factual Associations in GPT*, arXiv:2202.05262 (2022).

[134] Katie Atkinson, Trevor Bench-Capon & Danushka Bollegala, *Explanation in AI and Law: Past, Present and Future*, Artificial Intelligence 289 (2020) at 1 ("insights from AI and Law, where explanation has long been a concern, may provide useful pointers for future development of explainable AI.").

[135] *See, e.g.,* P. Jonathon Phillips et al., *Four Principles of Explainable Artificial Intelligence,* Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8312 (September 2021) https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf; Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, Columbia Law Review 119, no. 7 1851-1886 (2019).

[136] Understanding the models is also important for dealing with "inner alignment" problem(s) that may arise with more powerful AI systems because it could help uncover instances of AI models deceiving humans, and situations where AI models are optimizing for solving subproblems, that lead to issues unforeseen in the original problem specification; *see, e.g.,* Evan Hubinger et al., *Risks from learned optimization in advanced machine learning systems* (2019). For more on explainable AI in the alignment context, *see, e.g.,* Y. Belinkov & J. Glass, *Analysis methods in neural language processing: A survey*, In Transactions of the Association for Computational Linguistics, 7: 49–72 (2019); Evan Hubinger, *Relaxed adversarial training for inner alignment* (September 10, 2019) https://www.alignmentforum.org/posts/9Dy5YRaoCxH9zuJqa/relaxed-adversarial-training-for-inner-alignment. For more on interpreting the inner workings of AI models, *see, e.g.,* Chris Olah et al., *Zoom in: An introduction to circuits*, 5.3 e00024-001 Distill (March 10, 2020) https://distill.pub/2020/circuits/zoom-in/; Chelsea Voss et al., *Visualizing weights* 6.2 e00024-007 Distill (2021).

[137] *See, e.g.,* Chris Olah, *Mechanistic Interpretability, Variables, and the Importance of Interpretable Bases* (June 27, 2022) https://transformer-circuits.pub/2022/mech-interp-essay/index.html.

[138] This is possible because legal concepts are the ontology of our alignment framework and, in these cases where the deep learning models learn representations that can be identified as legal knowledge, legal concepts are a substrate of the model.

models throughout action-state space can uncover patterns of agent decision-making.[139] Safety benchmarks have been developed for simple environments for AI agents trained with reinforcement learning.[140] Similar approaches could help demonstrate an AI's comprehension of legal standards before the system is permitted to interact in the real word.[141] This would not be a fool-proof deterministic verification.[142] From a legal perspective, this is analogous to the certification of legal and regulatory understanding for professionals such as financial advisors, with the key difference that there is a relatively (computationally) costless assessment of AI legal understanding that is possible. Relative to the professional certification and subsequent testing we currently impose on humans providing specialized services such as financial advising, it is significantly less expensive to run millions of simulations of scenarios to test an AI's comprehension of relevant legal standards and regulations.[143]

---

[139] *See, e.g.,* John Nay, *A Machine Learning Approach to Modeling Dynamic Decision-Making in Strategic Interactions and Prediction Markets*, Vanderbilt University (2017) https://www.proquest.com/pagepdf/2007010189?accountid=12768; John Nay & Yevgeniy Vorobeychik, *Predicting Human Cooperation*, PloS One 11, no. 5 (2016).

[140] *See, e.g.,* Alex Ray, Joshua Achiam & Dario Amodei, *Benchmarking Safe Exploration in Deep Reinforcement Learning* (2019); Daniel S. Brown, Jordan Schneider, Anca Dragan & Scott Niekum, *Value Alignment Verification*, In International Conference on Machine Learning, 1105-1115. PMLR (2021).

[141] *See, e.g.,* Malgieri & Pasquale, *From Transparency to Justification: Toward Ex Ante Accountability for AI* (2022) (they propose "a system of "unlawfulness by default" for AI systems, an ex-ante model where some AI developers have the burden of proof to demonstrate that their technology is not discriminatory, not manipulative, not unfair, not inaccurate, and not illegitimate in its legal bases and purposes."). For other proposals related to the certification or verification of AI systems before and during deployment, *see, e.g.,* Miles Brundage et al., *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims* (2020); Inioluwa Deborah Raji et al., *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance* (2022); Gregory Falco et al., *Governing AI Safety Through Independent Audits*, Nature Machine Intelligence, Vol 3, 566–571 (2021); Peter Cihon et al., *AI Certification: Advancing Ethical Practice by Reducing Information Asymmetries*, IEEE Transactions on Technology and Society 2.4, 200-209 (2021); Andrew Tutt, *An FDA For Algorithms*, Admin. L. Rev. Vol. 69, No. 1, 83 (2017) at 122 (argues that there is a "close analog between complex pharmaceuticals and sophisticated algorithms[...]" and therefore the FDA provides a model for a new regulatory agency for algorithms); Florian Moslein & Roberto Zicari, *Certifying Artificial Intelligence Systems, in* BIG DATA LAW (Roland Vogl ed., 2021); Thomas Arnold & Matthias Scheutz, *The "big red button" is too late: an alternative model for the ethical evaluation of AI systems*, Ethics and Information Technology (2018) at 60 ("We outline a system architecture consisting of an ethical core layer above the hardware and below the virtual machine layer that consists of scenario-generation and simulation engines,"). For a proposal for verification of AI systems before deployment, specifically related to reinforcement learning decision-making systems, *see, e.g.,* Thomas Krendl Gilbert, Sarah Dean, Tom Zick & Nathan Lambert, *Choices, Risks, and Reward Reports: Charting Public Policy for Reinforcement Learning Systems* (2022).

[142] For a discussion of the difficulty of verifying the security of software systems and the analogy to aligning AI, *see, e.g.,* elspood, *Security Mindset: Lessons from 20+ years of Software Security Failures Relevant to AGI Alignment* (June 21, 2022).

[143] For a proposal on monitoring AI systems once they are deployed, *see, e.g.,* Amitai Etzioni & Oren Etzioni, *Keeping AI Legal*, 19 Vanderbilt Journal of Entertainment and Technology Law 133 (2016) at 139, 146 ("From here on, AI should be divided into two categories. The first category would consist of operational AI programs - the computerized "brains" that guide smart instruments. The second category would be composed of oversight AI programs that review the first category's decision making and keep the decisions in line with the law. These oversight programs, which this Article calls "AI Guardians," would include AI programs to interrogate, discover, supervise, audit, and guarantee the compliance of operational AI programs. [...] These AI Guardians will need to become smarter just as operational Al programs are improving [...] humans may have little choice but to draw on AI to check AI - and to seek to increase oversight of artificial intelligence as the intelligence of the programs they oversee grows.").

3. Legal Understanding as The Alignment Benchmark

Progress in AI research is driven, in large part, by shared benchmarks that thousands of researchers globally use to guide their experiments, understand as a community whether certain model and data advancements are adding value, and compare results across research groups.[144] Optimizing benchmarks are one of the primary "objective functions" of the overall global AI research apparatus.[145] As quantitative lodestars, benchmarks also create perverse incentives to build AI systems that optimize for benchmark performance at the expense of true generalization and intelligence,[146] see, e.g., "Goodhart's Law," colloquially communicated as, "when a measure becomes a target, it ceases to be a good measure."[147] Many AI benchmark datasets have a significant number of errors,[148] which suggests that, in some cases, machine learning models have, more than widely recognized, "overfitted to memorizing data instead of learning abstract concepts."[149] There are spurious cues within benchmark data structures that, once removed, significantly drop model performance, demonstrating that models are often learning patterns that do not generalize outside of the closed world of the benchmark data.[150] Many benchmarks, especially in natural language processing, have become quickly saturated,[151] as "contemporary models quickly achieve outstanding performance on benchmark tasks but nonetheless fail on simple challenge examples and falter in real-world scenarios."[152] Benchmarking AI capabilities is difficult. Benchmarking AI alignment has the same issues, but compounded by significantly vaguer problem definitions. There is also far less research on AI alignment benchmarks.

Performing well on societal alignment is more difficult than performing well on task capabilities.[153] Because alignment is so fundamentally hard, the sky's the limit on the difficulty of alignment benchmarks.[154] Legal-informatics-based benchmarks could serve as AI alignment

---

[144] Douwe Kiela et al., *Dynabench: Rethinking Benchmarking in NLP*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 4110–4124 (Association for Computational Linguistics, 2021) [Hereinafter Douwe Kiela, *Dynabench*].

[145] For more on this generalized notion of an objective function, *see*, Kenneth O. Stanley & Joel Lehman, *Why Greatness Cannot Be Planned: The Myth of the Objective* (2015).

[146] *See, generally*, the footnote above discussing the shortcut rule (François Chollet, *Deep Learning with Python, Second Edition* (2021) at 450). *See, e.g.,* Adriano Barbosa-Silva et al., *Mapping global dynamics of benchmark creation and saturation in artificial intelligence* (2022) at 1 ("We curated data for 1688 benchmarks covering the entire domains of computer vision and natural language processing, and show that a large fraction of benchmarks quickly trended towards near-saturation.").

[147] Charles Goodhart, *Problems of Monetary Management: The U.K. Experience,* In Courakis, Anthony S. (ed.). Inflation, Depression, and Economic Policy in the West (1975); *See, e.g.,* Peter Coy, *Goodhart's Law Rules the Modern World. Here Are Nine Examples*, Bloomberg.com (March 26, 2021); David Manheim & Scott Garrabrant, *Categorizing Variants of Goodhart's Law* (2018).

[148] Curtis G. Northcutt, Anish Athalye & Jonas Mueller, *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks* (2021).

[149] Björn Barz & Joachim Denzler, *Do We Train on Test Data? Purging CIFAR of Near-Duplicates* (2020) at 1.

[150] *See, e.g.,* Ronan Le Bras et al., *Adversarial Filters of Dataset Biases*, in Proceedings of the 37th International Conference on Machine Learning, PMLR 119:1078-1088 (2020).

[151] Samuel Bowman & George Dahl, *What Will it Take to Fix Benchmarking in Natural Language Understanding?*, in Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (2021).

[152] Douwe Kiela *Dynabench* at 4110.

[153] *See, e.g.,* the human *in*ability to align large numbers of humans or different groups, relative to extraordinary human capabilities on tasks in isolation or small groups.

[154] Researchers are making progress on better characterizing performance on task capabilities beyond simple comparisons of overall performance on monolithic benchmarks, *see, e.g.,* Kawin Ethayarajh et al., *Information-*

benchmarks for the research community. Current machine learning models perform poorly on legal understanding tasks such as statutory reasoning[155] and legal discovery;[156] there is significant room for improvement on legal text natural language processing tasks.[157] An example benchmark that could be used as part of the alignment benchmarks is Law Search.[158]

C. Legal Processes, Data & Experts Can Improve AI

Legal informatics may allow us to engineer legal data (both observational data and data derived from human interaction with models) into training signals that help align AI. Furthermore, legal informatics may actually improve general AI capabilities, which is important for accomplishing the ultimate alignment goal because techniques that increase AI alignment or safety at the expense of AI capabilities (the so-called "alignment tax"[159]) can lead to organizations eschewing alignment to gain additional capabilities[160] as organizations and countries race forward on developing and deploying AI.[161] If the safer version of AI performs better, then it is more likely to be adopted.

1. Models

Theoretic Measures of Dataset Difficulty (2021); Kawin Ethayarajh et al., *Understanding Dataset Difficulty with V-Usable Information* (2022). This work could inform methods to accurately measure alignment capabilities as well.

[155] *See*, recent applications in tax law, which is a more rules-based (as opposed to standards-based) area of the U.S. Code, *e.g.*, Nils Holzenberger, Andrew Blair-Stanek & Benjamin Van Durme, *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering* (2020); Nils Holzenberger & Benjamin Van Durme, *Factoring Statutory Reasoning as Language Understanding Challenges* (2021).

[156] Eugene Yang et al., *Goldilocks: Just-Right Tuning of BERT for Technology-Assisted Review*, in Advances in Information Retrieval: 44th European Conference on IR Research, 502–517 (Apr 2022).

[157] *See, e.g.*, Ilias Chalkidis et al., *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*, *in* PROCEEDINGS OF THE 60TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2022); D. Jain, M.D. Borah & A. Biswas, *Summarization of legal documents: where are we now and the way forward*, Comput. Sci. Rev. 40, 100388 (2021).

[158] *See*, Faraz Dadgostari et al., *Modeling Law Search as Prediction*, A.I. & L. 29.1, 3-34 (2021) at 3 ("In any given matter, before legal reasoning can take place, the reasoning agent must first engage in a task of "law search" to identify the legal knowledge—cases, statutes, or regulations—that bear on the questions being addressed."); Michael A. Livermore & Daniel N. Rockmore, *The Law Search Turing Test*, in Law as Data: Computation, Text, and the Future of Legal Analysis (2019) at 443-452; Michael A. Livermore et al., *Law Search in the Age of the Algorithm*, Mich. St. L. Rev. 1183 (2020).

[159] Askell *Laboratory for Alignment*.

[160] Eliezer Yudkowsky, *Aligning an AGI adds significant development time* (Feb 21, 2017) https://arbital.com/p/aligning_adds_time/; Askell *Laboratory for Alignment* ("Controlling the inputs and capabilities of AI systems will clearly have costs, so it seems hard to ensure that these controls, even if they're developed, are actually used."); Tom Adamczewski, *A Shift in Arguments for AI Risk* (May 25, 2019) at Section 3.2 https://fragile-credences.github.io/prioritising-ai/#the-importance-of-competitive-pressures.

[161] *See*, *e.g.*, Henry Kissinger, Eric Schmidt & Daniel P. Huttenlocher, *The Age of AI: And Our Human Future* (2021); Amanda Askell, Miles Brundage & Gillian Hadfield, *The Role of Cooperation in Responsible AI Development*, arXiv:1907.04534 (2019); Stephen Cave & Seán S. ÓhÉigeartaigh, *An AI Race for Strategic Advantage: Rhetoric and Risks*, In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 36-40 (2018).

The 2017-2021 era of deep learning AI advancements was, in large part, characterized by massively scaling the size of the models.[162] Law can better inform code by implementing these advancements in the legal informatics realm.[163] There now seems to be evidence that the bottleneck in pushing AI capability performance further in large language models[164] and recommendation models[165] is less on the sheer size of the models and more on the amount of useful data, training procedures, and possibly inductive biases (model architectures).[166] Legal informatics is an untapped source of data, and inspiration for training procedures and model structures.[167]

i. AI Capabilities Can Improve Legal Informatics Alignment

The *Law Informs Code* agenda can leverage recent advancements in machine learning,[168] in particular, natural language processing[169] with large language models trained with self-supervision; deep reinforcement learning; the intersection of large language models and deep reinforcement learning;[170] and research on "safe reinforcement learning"[171] (especially where constraints on agent actions can be described in natural language[172]). The combination of *(a)* large

---

[162] Computing power is a key factor in AI model performance, *see, e.g.*, Richard Sutton, *The Bitter Lesson*, Incomplete Ideas (blog), 13 12 (2019) http://www.incompleteideas.net/IncIdeas/BitterLesson.html. ("The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin. The ultimate reason for this is Moore's law, or rather its generalization of continued exponentially falling cost per unit of computation.").

[163] *See infra* Section II.C.1.i.

[164] Jordan Hoffmann et al., *Training Compute-Optimal Large Language Models* (2022) https://arxiv.org/abs/2203.15556.

[165] Newsha Ardalani et al., *Understanding Scaling Laws for Recommendation Models* (2022) https://arxiv.org/abs/2208.08489.

[166] Marcos Treviso et al., *Efficient Methods for Natural Language Processing: A Survey* (Aug 31, 2022) https://arxiv.org/abs/2209.00099.

[167] *See infra* Section II.C.1.ii.

[168] *See, e.g.*, Daniel Martin Katz & John Nay, *Machine Learning and Law*, *in* LEGAL INFORMATICS (Daniel Martin Katz et al. eds., 2021).

[169] For natural language processing methods applied to legal text, *see, e.g.*, John Nay, *Natural Language Processing for Legal Texts*, *in* LEGAL INFORMATICS (Daniel Martin Katz et al. eds. 2021); Michael A. Livermore & Daniel N. Rockmore, *Distant Reading the Law*, in Law as Data: Computation, Text, and the Future of Legal Analysis (2019) 3-19; J.B. Ruhl, John Nay & Jonathan Gilligan, *Topic Modeling the President: Conventional and Computational Methods*, 86 GEO. WASH. L. REV. 1243 (2018); John Nay, *Predicting and Understanding Law-making with Word Vectors and an Ensemble Model*, 12 PLOS ONE 1 (2017); John Nay, *Gov2Vec: Learning Distributed Representations of Institutions and Their Legal Text*, in Proceedings of 2016 Empirical Methods in Natural Language Processing Workshop on NLP and Computational Social Science, 49–54, Association for Computational Linguistics, (November 5, 2016).

[170] *See, e.g.,* Prithviraj Ammanabrolu et al., *Aligning to Social Norms and Values in Interactive Narratives* (2022).

[171] *See, generally,* Javier Garcia & Fernando Fernandez, *A Comprehensive Survey on Safe Reinforcement Learning*, Journal of Machine Learning Research, 16, 1 (2015) at 1437 ("Safe Reinforcement Learning can be defined as the process of learning policies that maximize the expectation of the return in problems in which it is important to ensure reasonable system performance and/or respect safety constraints during the learning and/or deployment processes."). *See, e.g.,* Philip S. Thomas et al., *Preventing undesirable behavior of intelligent machines*, Science 366.6468 999-1004 (2019); William Saunders et al., *Trial without Error: Towards Safe Reinforcement Learning via Human Intervention* (2017).

[172] *See, e.g.,* Tsung-Yen Yang et al., *Safe Reinforcement Learning with Natural Language Constraints* (2021) at 3, 2 (Most research on safe reinforcement learning requires "a human to specify the cost constraints in mathematical or logical form, and the learned constraints cannot be easily reused for new learning tasks. In this work, we design a

language models trained on large corpora of (sometimes morally salient[173]) text powering decision-making agents,[174] with *(b)* procedures that learn an automated mapping from natural language to environment dynamics[175] and reward functions of those agents,[176] and *(c)* offline

---

modular architecture to learn to interpret textual constraints, and demonstrate transfer to new learning tasks." Tsung-Yen Yang et al. developed "*Policy Optimization with Language COnstraints* (POLCO), where we disentangle the representation learning for textual constraints from policy learning. Our model first uses a *constraint interpreter* to encode language constraints into representations of forbidden states. Next, a *policy network* operates on these representations and state observations to produce actions. Factorizing the model in this manner allows the agent to retain its constraint comprehension capabilities while modifying its policy network to learn new tasks. Our experiments demonstrate that our approach achieves higher rewards (up to 11x) while maintaining lower constraint violations (up to 1.8x) compared to the baselines on two different domains."); Bharat Prakash et al., *Guiding safe reinforcement learning policies using structured language constraints*, UMBC Student Collection (2020).

[173] *See, e.g.,* Jin et al., *When to Make Exceptions: Exploring Language Models as Accounts of Human Moral Judgment* (2022) https://drive.google.com/file/d/1hB_yStLu52sGUYOVwq1aQ8O-6IeqR12g/view and https://github.com/feradauto/MoralCoT ("we present a novel challenge set consisting of rule-breaking question answering (RBQA) of cases that involve potentially permissible rule-breaking – inspired by recent moral psychology studies. Using a state-of-the-art large language model (LLM) as a basis, we propose a novel moral chain of thought (MORALCOT) prompting strategy that combines the strengths of LLMs with theories of moral reasoning developed in cognitive science to predict human moral judgments."); Liwei Jiang et al., *Delphi: Towards Machine Ethics and Norms* (2021) ("1.7M examples of people's ethical judgments on a broad spectrum of everyday situations"); Dan Hendrycks et al., *Aligning AI With Shared Human Values* (2021) at 2 ("We find that existing natural language processing models pre-trained on vast text corpora and fine- tuned on the ETHICS dataset have low but promising performance. This suggests that current models have much to learn about the morally salient features in the world, but also that it is feasible to make progress on this problem today."); Frazier et al., *Learning Norms from Stories: A Prior for Value Aligned Agents* (2019).

[174] *See, e.g.,* Prithviraj Ammanabrolu et al., *Aligning to Social Norms and Values in Interactive Narratives* (2022) ("We introduce [...] an agent that uses the social commonsense knowledge present in specially trained language models to contextually restrict its action space to only those actions that are aligned with socially beneficial values."); Md Sultan Al Nahian et al., *Training Value-Aligned Reinforcement Learning Agents Using a Normative Prior* (2021) ("We introduce an approach to value-aligned reinforcement learning, in which we train an agent with two reward signals: a standard task performance reward, plus a normative behavior reward. The normative behavior reward is derived from a value-aligned prior model previously shown to classify text as normative or non-normative. We show how variations on a policy shaping technique can balance these two sources of reward and produce policies that are both effective and perceived as being more normative."); Dan Hendrycks et al., *What Would Jiminy Cricket Do? Towards Agents That Behave Morally* (2021) ("To facilitate the development of agents that avoid causing wanton harm, we introduce Jiminy Cricket, an environment suite of 25 text-based adventure games with thousands of diverse, morally salient scenarios. By annotating every possible game state, the Jiminy Cricket environments robustly evaluate whether agents can act morally while maximizing reward."); Shunyu Yao et al., *Keep CALM and Explore: Language Models for Action Generation in Text-based Games* (2020) ("Our key insight is to train language models on human gameplay, where people demonstrate linguistic priors and a general game sense for promising actions conditioned on game history. We combine CALM with a reinforcement learning agent which re-ranks the generated action candidates"); Matthew Hausknecht et al., *Interactive Fiction Games: A Colossal Adventure* (2019) at 1 ("From a machine learning perspective, Interactive Fiction games exist at the intersection of natural language processing and sequential decision making. Like many NLP tasks, they require natural language understanding, but unlike most NLP tasks, IF games are sequential decision making problems in which actions change the subsequent world states").

[175] *See e.g.,* Austin W. Hanjie, Victor Zhong & Karthik Narasimhan, *Grounding Language to Entities and Dynamics for Generalization in Reinforcement Learning* (2021); Prithviraj Ammanabrolu & Mark Riedl, *Learning Knowledge Graph-based World Models of Textual Environments*, In Advances in Neural Information Processing Systems 34 3720-3731 (2021); Felix Hill et al., *Grounded Language Learning Fast and Slow* (2020); Marc-Alexandre Côté et al., *TextWorld: A Learning Environment for Text-based Games* (2018).

[176] *See e.g.,* MacGlashan et al., *Grounding English Commands to Reward Functions, Robotics: Science and Systems* (2015) at 1 ("Because language is grounded to reward functions, rather than explicit actions that the robot can perform, commands can be high-level, carried out in novel environments autonomously, and even transferred to other robots with different action spaces. We demonstrate that our learned model can be both generalized to novel environments

reinforcement learning (especially with Transformer-based models)[177] represents a potential opportunity to leverage millions (or even billions) of state-action-value tuples from (natural language) legal text within reinforcement or imitation learning paradigms (where AI agents make "decisions").

Further research should experiment with how legal informatics is employed within AI agent decision-making paradigms,[178] e.g., *(a)* as (natural language) constraints in the optimization problem,[179] *(b)* for shaping the reward function during training,[180] *(c)* for refined representations of the state space,[181] *(d)* for guiding the exploration of the state space during training,[182] *(e)* as inputs to world models for efficient training,[183] *(f)* as a Foundation Model prior, or part of pretraining, to bias a deployed agent's action space toward certain actions or away from others,[184] or *(g)* as some combination of the aforementioned. Where legal informatics is providing the modular constructs (methods of statutory interpretation, applications of modularized standards, and legal reasoning more broadly) to facilitate the communication of what a human wants an AI system to do, it is more likely that it is employed for specifying and shaping reward functions. Where legal informatics, through public law and policy, helps specify what AI systems should *not* do, in order to provide a broader knowledge base of how to reduce societal externalities, it is more likely employed as action constraints.

"Legal decision-making requires context at various scales: knowledge of all historical decisions and standards, knowledge of the case law that remains relevant in the present, and knowledge of the nuances of the individual case at hand. Foundation models are uniquely poised

---

and transferred to a robot with a different action space than the action space used during training."); Karthik Narasimhan, Regina Barzilay & Tommi Jaakkola, *Grounding Language for Transfer in Deep Reinforcement Learning* (2018); Prasoon Goyal, Scott Niekum & Raymond J. Mooney, *Using Natural Language for Reward Shaping in Reinforcement Learning* (2019); Jelena Luketina et al., *A Survey of Reinforcement Learning Informed by Natural Language* (2019); Theodore Sumers et al., *Learning Rewards from Linguistic Feedback* (2021); Jessy Lin et al., *Inferring Rewards from Language in Context* (2022); Pratyusha Sharma et al., *Correcting Robot Plans with Natural Language Feedback* (2022).

[177] *See, e.g.*, Lili Chen et al., *Decision Transformer: Reinforcement Learning via Sequence Modeling* (2021); Sergey Levine et al., *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems* (2020).

[178] *See, e.g.*, Mykel J. Kochenderfer, Tim A. Wheeler & Kyle H. Wray, *Algorithms for Decision Making*, MIT Press (2022).

[179] *See, e.g.*, Tsung-Yen Yang et al., *Safe Reinforcement Learning with Natural Language Constraints* (2021) at 3 ("Since constraints are decoupled from rewards and policies, agents trained to understand certain constraints can transfer their understanding to respect these constraints in new tasks, even when the new optimal policy is drastically different.").

[180] *See, e.g.*, Bharat Prakash et al., *Guiding safe reinforcement learning policies using structured language constraints*, UMBC Student Collection (2020); Dan Hendrycks et al., *What Would Jiminy Cricket Do? Towards Agents That Behave Morally* (2021).

[181] *See, e.g.*, Mengjiao Yang & Ofir Nachum, *Representation Matters: Offline Pretraining for Sequential Decision Making*, In Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 11784-11794 (2021).

[182] *See, e.g.*, Allison C. Tam et al., *Semantic Exploration from Language Abstractions and Pretrained Representations* (2022).

[183] *See, e.g.*, Vincent Micheli, Eloi Alonso & François Fleuret, *Transformers are Sample Efficient World Models* (2022) https://arxiv.org/abs/2209.00588.

[184] *See, e.g.*, Jacob Andreas, Dan Klein & Sergey Levine. *Learning with Latent Language*, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol 1, 2166–2179, Association for Computational Linguistics (2018)*;* Shunyu Yao et al., *Keep CALM and Explore: Language Models for Action Generation in Text-based Games* (2020); Andrew K Lampinen et al., *Tell me why! Explanations support learning relational and causal structure*, in Proceedings of the 39th International Conference on Machine Learning, PMLR 162, 11868-11890 (2022).

to have the potential to learn shared representations of historical and legal contexts, as well as have the linguistic power and precision for modeling an individual case."[185] Foundation Models trained on legal text have learned model weights and word embeddings specific to legal text that provide (slightly) better performance on downstream legal tasks[186] and have been useful for analyzing legal language,[187] and legal arguments.[188] Foundation Models' recent strong capabilities in automatically analyzing (non-legal) citations[189] may prove fruitful in identifying relevant legal precedent, and their ability to generate persuasive language could help AI understand, and thus learn from, legal brief text data.[190]

In many cases, Foundation Models are not truthful,[191] but they have become capable of more truthfulness as they are scaled.[192] Foundation Models are beginning to demonstrate improved performance in analyzing contracts.[193] As state-of-the-art models have gotten larger and more

---

[185] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) https://arxiv.org/pdf/2108.07258.pdf at 63.

[186] *See, e.g.*, Zheng et al., *When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings*, In ICAIL '21: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (June 2021), at 159 ("Our findings [...] show that Transformer-based architectures, too, learn embeddings suggestive of distinct legal language.").

[187] *See, e.g.*, Julian Nyarko & Sarath Sanga, *A Statistical Test for Legal Interpretation: Theory and Applications*, The Journal of Law, Economics, and Organization, https://doi.org/10.1093/jleo/ewab038 (2020); Jonathan H. Choi, *An Empirical Study of Statutory Interpretation in Tax Law*, NYU L Rev. 95, 363 (2020) https://www.nyulawreview.org/issues/volume-95-number-2/an-empirical-study-of-statutory-interpretation-in-tax-law/.

[188] Prakash Poudyal et al., *ECHR: Legal Corpus for Argument Mining*, In *Proceedings of the 7th Workshop on Argument Mining*, 67–75, Association for Computational Linguistics (2020) at 1 ("The results suggest the usefulness of pre-trained language models based on deep neural network architectures in argument mining.").

[189] *See, e.g.*, Petroni et al., *Improving Wikipedia Verifiability with AI* (2022) https://openreview.net/forum?id=qfTqRtkDbWZ.

[190] *See e.g.*, Sebastian Duerr & Peter A. Gloor, *Persuasive Natural Language Generation – A Literature Review* (2021); Jialu Li, Esin Durmus & Claire Cardie, *Exploring the Role of Argument Structure in Online Debate Persuasion*, In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 8905–8912 (2020); Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) https://arxiv.org/pdf/2108.07258.pdf at 64.

[191] Owain Evans, et al., *Truthful AI: Developing and Governing AI That Does Not Lie*, arXiv:2110.06674 (2021); S. Lin, J. Hilton & O. Evans, *TruthfulQA: Measuring How Models Mimic Human Falsehoods*, arXiv:2109.07958 (2021) http://arxiv.org/abs/2109.07958.

[192] *See e.g.*, Dan Hendrycks et al., *Measuring Massive Multitask Language Understanding*, arXiv:2009.03300 (2020); Jared Kaplan et al., *Scaling Laws for Neural Language Models*, arXiv:2001.08361 (2020). However, scaling, alone, is unlikely to make the models "fully truthful," *see,* Owain Evans, et al., *Truthful AI: Developing and Governing AI That Does Not Lie*, arXiv:2110.06674 (2021) at 61 (The data the models are trained on "contains many instances of humans being non-truthful and so [the model] will likely be non-truthful in the same contexts. In summary, we have a speculative argument that language modelling (without tweaks or modifications) is unlikely to produce truthful AI systems.").

[193] Spyretta Leivaditi, Julien Rossi & Evangelos Kanoulas, *A Benchmark for Lease Contract Review*, arXiv:2010.10386 (2020); Allison Hegel et al., *The Law of Large Documents: Understanding the Structure of Legal Contracts Using Visual Cues*, arXiv:2107.08128 (2021); Dan Hendrycks, Collin Burns, Anya Chen & Spencer Ball, *Cuad: An expert-annotated NLP dataset for legal contract review*, arXiv:2103.06268 (2021); Ilias Chalkidis et al., *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*, *in* PROCEEDINGS OF THE 60TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2022); Stephen C. Mouritsen, *Contract Interpretation with Corpus Linguistics*, 94 WASH. L. REV. 1337 (2019); Yonathan A. Arbel & Shmuel I. Becher, *Contracts in the Age of Smart Readers*, 83 Geo. Wash. L. Rev. 90 (2022).

advanced over time, their contract analysis performance has improved,[194] suggesting we can expect continued advancements in natural language processing capabilities to improve legal text analysis as a by-product.[195] These examples represent the preferred approach: *Law* can best *Inform Code* if legal informatics is able to convert the progress in general AI capabilities (which is being aggressively funded by most of the large internet technology companies and national governments) into gains in AI legal understanding.

Mainstream "capabilities" AI research focus areas could potentially unlock further advances toward *Law Informing Code* in the near-term. For instance, the successful application of deep reinforcement learning further beyond toy problems (e.g., video games and board games),[196] with human feedback,[197] and through offline learning at large scale.[198]

ii. Legal Informatics Alignment Could Improve AI Capabilities

This may eventually be a two-way street, with the legal informatics alignment research also improving fundamental AI capabilities.[199] We provide four example avenues.

---

[194] Dan Hendrycks, Collin Burns, Anya Chen & Spencer Ball, *Cuad: An expert-annotated NLP dataset for legal contract review*, arXiv:2103.06268 (2021) at 2 ("We experiment with several state-of-the-art Transformer (Vaswani et al., 2017) models on CUAD [a dataset for legal contract review]. We find that performance metrics such as Precision @ 80% Recall are improving quickly as models improve, such that a BERT model from 2018 attains 8.2% while a DeBERTa model from 2021 attains 44.0%.").

[195] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) https://arxiv.org/pdf/2108.07258.pdf at 59 ("Many legal applications pose unique challenges to computational solutions. Legal language is specialized and legal outcomes often rely on the application of ambiguous and unclear standards to varied and previously unseen fact patterns. At the same time, due to its high costs, labeled training data is scarce. Depending on the specific task, these idiosyncrasies can pose insurmountable obstacles to the successful deployment of traditional models. In contrast, their flexibility and capability to learn from few examples suggest that foundation models could be uniquely positioned to address the aforementioned challenges.").

[196] In the legal understanding domain, *see, e.g.,* Duy-Hung Nguyen et al., *Robust Deep Reinforcement Learning for Extractive Legal Summarization*, in International Conference on Neural Information Processing (2021).

[197] *See, e.g.,* Paul F. Christiano et al., *Deep Reinforcement Learning from Human Preferences,* in Advances in Neural Information Processing Systems 30 (2017); Natasha Jaques et al., *Way Off-Policy Batch Deep Reinforcement Learning of Implicit Human Preferences in Dialog* (2019); Stiennon et al., *Learning to Summarize with Human Feedback,* In 33 Advances in Neural Information Processing Systems 3008-3021 (2020); Daniel M. Ziegler, et al., *Fine-tuning Language Models From Human Preferences*, arXiv:1909.08593 (2019); Cassidy Laidlaw & Stuart Russell, *Uncertain Decisions Facilitate Better Preference Learning* (2021); Koster et al., *Human-centred mechanism design with Democratic AI,* Nature Human Behaviour (2022); Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*, arxiv.org (Mar. 4, 2022), https://arxiv.org/pdf/2203.02155.pdf.

[198] *See, e.g.,* Dibya Ghosh et al., *Offline RL Policies Should be Trained to be Adaptive* (2022); Machel Reid, Yutaro Yamada & Shixiang Shane Gu, *Can Wikipedia Help Offline Reinforcement Learning?* (2022); Sergey Levine et al., *Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems* (2020) at 25 (Combining offline and online RL through historical legal information and human feedback is likely a promising integrated approach, because, "if the dataset state-action distribution is narrow, neural network training may only provide brittle, non-generalizable solutions. Unlike online reinforcement learning, where accidental overestimation errors arising due to function approximation can be corrected via active data collection, these errors cumulatively build up and affect future iterates in an offline RL setting.").

[199] Rishi Bommasani et al., *On the Opportunities and Risks of Foundation Models*, arxiv.org (Aug. 18, 2021) https://arxiv.org/pdf/2108.07258.pdf at 66 ("Legal briefing and reasoning is likely beyond the capabilities of current models, but appears to be within the future realm of possibilities. As such, these serve as a potential lode star for the ongoing development of foundation models."); Bart Verheij, *Artificial Intelligence as Law,* A.I. & L. 28, no. 2, 181-206 (2020).

First, legal informatics alignment research could improve fundamental AI capabilities by inspiring novel inductive biases for deep learning models from legal analogical reasoning.[200] Neuro-symbolic modeling (combining parametric models such as deep neural networks with non-parametric symbolic systems) is a promising approach to building generalizable reasoning capabilities,[201] and legal informatics research can help with "the much needed step we have to make towards hybrid systems that connect knowledge representation and reasoning techniques with the powers of machine learning."[202]

Second, machine-readable-law research[203] and practice[204] seeks hybrid structured-unstructured data representations of legal directives. This work advances capabilities for specifying human objectives and directives (reward functions) in code, and could directly influence AI research, if framed appropriately.

Third, adversarial debate is fundamental to legal processes and there is a promising line of AI research – inspired by the success of self-play AI systems like AlphaGo[205] – that is pursuing the modeling of artificial debate between AI agents as a means of performing more advanced tasks than humans while remaining aligned with those humans.[206] This approach tackles the machine learning problem, "as a game played between two agents, where the agents have an argument with each other and the human judges the exchange. Even if the agents have a more advanced understanding of the problem than the human, the human may be able to judge which agent has the better argument (similar to expert witnesses arguing to convince a jury)."[207] Legal informatics

---

[200] For the importance of abstraction, and ideas for building it into machine learning models, *see, e.g.*, Murray Shanahan & Melanie Mitchell, *Abstraction for Deep Reinforcement Learning*, in Proceedings of the International Joint Conference on Artificial Intelligence (2022); Melanie Mitchell, *Abstraction and Analogy-Making in Artificial Intelligence*, in Annals of the New York Academy of Sciences, 1505 (1), 79–101 (2021) at 1 ("While AI has made dramatic progress over the last decade in areas such as computer vision, natural language processing, and robotics, current AI systems almost entirely lack the ability to form humanlike concepts and abstractions."). Analogies and metaphors are foundational to the law, *see e.g.*, Steven L. Winter, *The Metaphor of Standing and the Problem of Self-governance*, Stan. L. Rev. 40, 1371 (1987); Steven L. Winter, *A Clearing in the Forest* (2001).

[201] *See e.g.*, Rajarshi Das, et al., *Case-based Reasoning for Natural Language Queries over Knowledge Bases*, in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 9594–9611, Association for Computational Linguistics (2021); F Ward, F Belardinelli & F Toni, *Argumentative Reward Learning: Reasoning About Human Preferences*, in HMCaT 2022 ICML (2022).

[202] Bart Verheij, *Artificial Intelligence as Law*, A.I. & L. 28, no. 2, 181-206 (2020) at 191.

[203] *See e.g.*, Megan Ma & Bryan Wilson, *The Legislative Recipe: Syntax for Machine-Readable Legislation*, 19 Nw. J. Tech. & Intell. Prop. 107 (2021); J. Mohun & A. Roberts, *Cracking the Code: Rulemaking for Humans and Machines*, OECD Working Papers on Public Governance, No. 42, (2020); *Cambridge Regulatory Genome*, https://www.jbs.cam.ac.uk/faculty-research/centres/regulatory-genome-project/about-rgp/#crg (2020); Tom Barraclough, Hamish Fraser & Curtis Barnes, *Legislation as Code for New Zealand: Opportunities, Risks, and Recommendations* (March 2021) http://www.nzlii.org/nz/journals/NZLFRRp/2021/3.html; Sarah B. Lawsky, *Form as Formalization*, 16 Ohio St. Tech. LJ 114 (2020); Sarah B. Lawsky, *A Logic for Statutes*, 21 Fla. Tax Rev. 60 (2017); Marcos A. Pertierra, Sarah B. Lawsky, Erik Hemberg & Una-May O'Reilly, *Towards Formalizing Statute Law as Default Logic through Automatic Semantic Parsing*, in ASAIL@ ICAIL (2017).

[204] *See e.g.*, *OpenFisca*, https://openfisca.org/en/; *VisiLaw*, https://www.visilaw.com/; *Catala*, https://catala-lang.org/.

[205] The first program to defeat a professional human Go player, *see*, Silver et al., *Mastering the game of Go without human knowledge*, Nature, 550, 354–359 (2017).

[206] *See e.g.*, Richard Ngo, *Why I'm Excited About Debate* (2021) https://www.alignmentforum.org/posts/LDsSqXf9Dpu3J3gHD/why-i-m-excited-about-debate; Geoffrey Irving, Paul Christiano & Dario Amodei, *AI Safety Via Debate* (2018) https://arxiv.org/abs/1805.00899.

[207] The authors' "hope is that, properly trained, such agents can produce value-aligned behavior far beyond the capabilities of the human judge. If the two agents disagree on the truth but the full reasoning is too large to show the humans, the debate can focus in on simpler and simpler factual disputes, eventually reaching a claim that is simple enough for direct judging." Geoffrey Irving & Dario Amodei, *AI Safety via Debate* (May 3, 2018).

and computational argumentation modeling,[208] could inform this research.[209] For more advanced AI systems based on Foundation Models, there may be ways to leverage the way in which much of the legal process is inherently adversarial to embed capabilities for fending off adversarial attacks and for improving model training with adversarial and argumentation techniques. Securing machine learning systems against adversarial attacks is difficult;[210] simple systems can fail in unexpected and surprising ways, and more advanced systems can in some cases be easily fooled.[211] Approaches to adversarial training of machine learning models for improving (a) general task capabilities[212] and (b) human-AI intent alignment,[213] and adversarial benchmarking of machine learning models[214] may benefit.

Fourth, the entire *Law Informs Code* research agenda related to contracts is aimed at improving the ability of AI systems to more efficiently understand which actions to perform for a human, which is a core ability toward improving general AI capabilities.[215]

---

[208] Kevin D. Ashley & VR Walker, *Toward Constructing Evidence-based Legal Arguments Using Legal Decision Documents and Machine Learning*, in Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 176–180 ACM (2013); Katie Atkinson et al., *Toward Artificial Argumentation*, AI Mag 38(3):25–36 (2017); Bart Verheij, *Proof with and without probabilities. Correct evidential reasoning with presumptive arguments, coherent hypotheses and degrees of uncertainty*. A.I. & L. 25(1):127–154 (2017); P Baroni, D Gabbay, M Giacomin & van der Torre L (eds.) *Handbook of Formal Argumentation* (2018); Bart Verheij, *Artificial Intelligence as Law*, A.I. & L. 28, no. 2, 181-206 (2020) at 191-193.

[209] And, back in the other direction, the results of this research area can be used to further improve legal informatics, specifically, to leverage improved machine learning capabilities for simulating theoretical court outcomes, in order to draw provisional conclusions about the way in which a legal precedent or legal standard may apply to a particular circumstance.

[210] *See e.g.*, in the context of natural language processing, Linyang Li et al., *BERT-ATTACK: Adversarial Attack Against BERT Using BERT*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (2020); Eric Wallace et al., *Universal Adversarial Triggers for Attacking and Analyzing NLP*, in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (2019). *See e.g.*, in the context of deep reinforcement learning, Adam Gleave et al., *Adversarial Policies: Attacking Deep Reinforcement Learning*, In Proc. ICLR-20 (2020).

[211] Ian Goodfellow et al., *Attacking Machine Learning with Adversarial Examples* (February 24, 2017); Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples* (2015).

[212] *See, generally*, Anthony Huang et al., *Adversarial Machine Learning*, In Proceedings of the 4th ACM workshop on Security and Artificial Intelligence, 43–58 (2011). *See, e.g.*, Mingyang Yi et al., *Improved OOD Generalization via Adversarial Training and Pretraining*, In Proceedings of the 38th International Conference on Machine Learning, Vol 139 11987, 18–24 (Jul 2021) https://proceedings.mlr.press/v139/yi21a.html; H. Zhang, I. Goodfellow, D. Metaxas & A. Odena, *Self-attention Generative Adversarial Networks*, in International Conference on Machine Learning, 7354-7363 (May 2019). Although, note that, as of 2022, diffusion models are generally a more capable approach than GANs, *see, e.g.*, Prafulla Dhariwal & Alexander Nichol, *Diffusion models beat GANs on image synthesis*, In Advances in Neural Information Processing Systems 34 8780 (2021); Robin Rombach et al., *High-resolution image synthesis with latent diffusion models*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2022).

[213] *See, e.g.*, Daniel M. Ziegler et al., *Adversarial Training for High-Stakes Reliability* (2022) https://arxiv.org/abs/2205.01663. Adversarial examples can be understood as "a misalignment between the (human-specified) notion of robustness and the inherent geometry of the data", *see*, Andrew Ilyas et al., *Adversarial Examples Are Not Bugs, They Are Features*, in Advances in neural information processing systems 32 (2019).

[214] *See, e.g.*, Yixin Nie et al., *Adversarial NLI: A New Benchmark for Natural Language Understanding*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (2020).

[215] *See infra* Section III.

2. Data

In addition to refining our theoretical understanding of the problem and guiding design of AI architectures, the legal informatics approach suggests sources of data for model training, fine-tuning, and validation.

i. Data from Experts

Evaluating a behavior is easier than learning how to actually execute that behavior; for example, I cannot do a backflip but I can evaluate whether you just did a backflip.[216] With this in mind, reinforcement learning through human attorney feedback (there are more than 1.3 million lawyers in the US[217]) on natural language interactions with AI models is potentially a powerful process to teach (through training, or fine-tuning, or extraction of templates for in-context prompting of large language models[218]) statutory interpretation, argumentation, and case-based reasoning, which can then be applied broadly for AI alignment. With large language models, only a few samples of human feedback, in the form of natural language, are needed for model refinement for some tasks.[219] Models could be trained to assist human attorney evaluators, which theoretically, in partnership with the humans, could allow the combined human-AI evaluation team to have capabilities that surpass the legal understanding of the expert humans alone.[220]

ii. Data from Legal Text

The Foundation Models in use today have effectively conducted a form of behavioral cloning on a large portion of the Internet to leverage billions of human actions (through natural language expressions). It may be possible to, similarly, leverage billions of human legal data points to build Law Foundation Models through large-scale language model self-supervision on pre-processed legal text data.[221]

---

[216] Jan Leike et al., *Scalable agent alignment via reward modeling: a research direction* (2018).
[217] *ABA Profile of the Legal Profession* (2022) at 22 https://www.americanbar.org/content/dam/aba/administrative/news/2022/07/profile-report-2022.pdf.
[218] *See, e.g.,* D. Khashabi, C. Baral, Y Choi & H Hajishirzi, *Reframing Instructional Prompts to GPTk's Language*, In Findings of the Association for Computational Linguistics, 589-612 (May 2022). Large neural network models have demonstrated the ability to learn mathematical functions purely from in-context interaction, *see, e.g.,* Shivam Garg, Dimitris Tsipras, Percy Liang & Gregory Valiant, *What Can Transformers Learn In-Context? A Case Study of Simple Function Classes* (2022).
[219] *See, e.g.,* Jérémy Scheurer et al., *Training Language Models with Language Feedback* (2022).
[220] See, e.g., William Saunders et al., *Self-critiquing models for assisting human evaluators* (2022).
[221] *See, e.g.,* Zheng et al., *When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings*, In ICAIL '21: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law (June 2021); Ilias Chalkidis et al., *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English*, *in* PROCEEDINGS OF THE 60TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS (2022); Ilias Chalkidis et al., *LEGAL-BERT: The Muppets straight out of Law School*, in Findings of the Association for Computational Linguistics: EMNLP, 2898-2904 (November 2020); Peter Henderson et al., *Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset* (2022) https://arxiv.org/abs/2207.00220.

Selecting which data sets are best suited for general self-supervised pre-training is an active area of research.[222] This is especially important in the legal domain where many historical actions represent institutionalized prejudices and politics.[223] We use two filters to guide data selection and data structuring processes. *First*, is the goal of training on a data point to embed world knowledge into AI or task knowledge. Learning that humans in the U.S. drive on the right side of the road is learning world knowledge; whereas, learning how to map a statute about driving rules to a new fact pattern in the real world is learning how to conduct a legal reasoning task. *Second*, is the approximate nature of the uncertainty as epistemic vs. aleatory[224] that an AI could theoretically resolve by training on a data point. If the nature of the uncertainty is epistemic – e.g., whether citizens prefer climate change risk reduction over endangered species protection – then it is fruitful to apply as much data as we can to learning functions to closer approximate the underlying fact about the world or about law. If the nature of the uncertainty is more of an aleatory flavor – e.g., the middle name of the defendant in a case – then there is enough inherent randomness that we would seek to avoid attempting to learn anything about that fact or data point.[225]

Legal standards can be learned directly from legal data.[226] We can also codify examples of human and corporate behavior exhibiting standards such as fiduciary duty into a structured format to evaluate the standards-understanding capabilities of AI models.[227] The legal data available for AI systems to learn from, or be evaluated on, includes textual data from all types of law (constitutional, statutory, administrative, case, and contractual),[228] legal training tools (e.g., bar exam outlines, casebooks, and software for teaching the casuistic approach), rule-based legal reasoning programs,[229] and human-in-the-loop live feedback from law and policy human

[222] *See, e.g.*, Thao Nguyen et al., *Quality Not Quantity: On the Interaction between Dataset Design and Robustness of CLIP* (2022).

[223] *See, e.g.*, Kathryn Stanchi, *The Rhetoric of Racism in the United States Supreme Court*, BCL Rev. 62, 1251 (2021).

[224] These are rough abstractions, and any determination of their application should be interpreted to be on a continuum, and itself highly uncertain.

[225] Discerning whether content is epistemic vs. aleatory is a major hurdle, and context dependent.

[226] *See, e.g.*, Peter Henderson et al., *Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset* (2022) at 7 https://arxiv.org/abs/2207.00220 (They learn data filtering standards related to privacy and toxicity from legal data, *e.g.*, "a model trained on Pile of Law (pol-bert) ranks Jane Doe ∼ 3 points higher than a standard bert-large-uncased on true pseudonym cases. This suggests that models pre-trained on Pile of Law are more likely to encode appropriate pseudonymity norms. To be sure, pol-bert is slightly more biased for Jane Doe use overall, as is to be expected, but its performance gains persist even after accounting for this bias.").

[227] This data could include both "gold-standard" human labeled data, but also automated data structuring that is subsequently sampled and selectively human validated for correctness. Data hand-labeled by expensive legal experts is unlikely to provide a large enough data set for training large neural models. Rather, its purpose is to validate the performance of models trained on much larger, general data, e.g., Foundation Models trained on large portions of the Internet. This semi-structured data could be used to design self-supervised learning processes to apply across relevant case law, regulatory guidance, training materials, and self-regulatory organization data (*e.g.*, FINRA exams) in order to train models to learn correct and incorrect fiduciary behavior across as many contexts as possible.

[228] In the U.S., the legislative branch creates statutory law through bills enacted by Congress; the executive branch creates administrative regulations through Agencies' notices of proposed rule-making and final rules; the judicial branch creates case law through judicial opinions; and private parties create contracts. Laws are found at varying levels of government in the United States: federal, state, and local. The adopted versions of public law are often compiled in official bulk data repositories that offer machine-readable formats. For instance, statutory law is integrated into the United States Code (or a state's Code), which organizes the text of all Public Laws that are still in force into subjects; and administrative policies become part of the Code of Federal Regulations (or a state's Code of Regulations), also organized by subject.

[229] *See, e.g.*, HYPO (Kevin D. Ashley, *Modelling Legal Argument: Reasoning With Cases and Hypotheticals* (1989)); CATO (Vincent Aleven, *Teaching Case-based Argumentation Through a Model and Examples* (1997)); Latifa Al-

experts.[230] The latter two could simulate state-action-reward spaces for AI fine-tuning or validation, and the former could be processed to do so.

Automated data curation processes to convert textual legal data into either state-action-reward tuples, or contextual constraints for shaping candidate action choices conditional on the state, is an important frontier in this research agenda (and promising for application to case law text data, contracts, and legal training materials). Learning from textual descriptions, rather than direct instruction, may allow models to learn reward functions that better generalize;[231] fortunately, much of law is embedded more in the form of descriptions and standards than it is in the form of direct instructions and specific rules. Descriptions of the application of standards provides a rich and large surface area to learn from.

Textual data can be curated and labeled for these purposes. Efforts of this nature should aim for two outcomes. First, data that can be used to evaluate how well AI models understand legal standards. Second, the possibility that the initial "gold-standard" human expert labeled data can be used to generate additional much larger sets of data through automated curation and processing of full corpora of legal text, and through model interaction with human feedback.

The first step is generating labeled data that can be used to evaluate the ability of an AI system to exhibit behavior consistent with particular legal standards. This can be broken into two types of tasks. First, a *prediction task*, e.g., "Is this piece of text about fiduciary standards?" Second, a *decision task*, e.g., "Here is the context of a scenario (the state of the world) [...], and a description of a choice to be made [...]. The action choices are [X]. The chosen action was [x]. This violates [legal standard z], [(if available) with [positive / negative] repercussions (reward)]." In other words, the "world" state (circumstance), potential actions, actual action, and the "reward" associated with taking that action in that state are labeled.[232]

Initial "gold-standard" human expert labeled data could eventually be used for training models to generate many more automated data labels. Models could be trained to map from the raw legal text to the more structured data outlined above. If successful, this phase of the legal

Abdulkarim, Katie Atkinson & Trevor Bench-Capon, *A Methodology for Designing Systems To Reason With Legal Cases Using Abstract Dialectical Frameworks*, A.I. & L. Vol 24, 1–49 (2016). *See, generally*, Katie Atkinson, *Representation and Automation of Legal Information*, *in* BIG DATA LAW (Roland Vogl ed., 2021).

[230] Research at the intersection of social, behavioral, and computer science is helping to determine which methods for learning from human feedback and human observational behavior are most effective in a reinforcement learning setting, *see, e.g.,* David Lindner & Mennatallah El-Assady, *Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning* (2022); Mihaela Curmei et al., *Towards Psychologically-Grounded Dynamic Preference Models*, arXiv 2208.01534 (2022).

[231] *See,* Sumers et al., *How To Talk So Your Robot Will Learn: Instructions, Descriptions, and Pragmatics* (2022); Karthik Narasimhan, Regina Barzilay & Tommi Jaakkola, *Grounding Language for Transfer in Deep Reinforcement Learning* (2018); Austin W. Hanjie, Victor Zhong & Karthik Narasimhan, *Grounding Language to Entities and Dynamics for Generalization in Reinforcement Learning* (2021).

[232] Like all machine learning models, natural language processing focused models often learn spurious associations, *see, e.g.,* Divyansh Kaushik & Zachary C. Lipton, *How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks,* in Empirical Methods in Natural Language Processing (2018). To address this, and learn more generalizable knowledge from textual data, it is helpful to obtain counterfactual label augmentations (*see,* Divyansh Kaushik, Eduard Hovy & Zachary C. Lipton, *Learning the Difference that Makes a Difference with Counterfactually-Augmented Data* (2020) (Labels that revise each textual input, "so that it (i) accords with a counterfactual target label; (ii) retains internal coherence; and (iii) avoids unnecessary changes"); Matt Gardner et al., *Evaluating Models' Local Decision Boundaries via Contrast Sets* (2020)), and to remove data labels where spurious artifacts are likely to lead the models to learn patterns that do not generalize; *see, e.g.,* Ronan Le Bras et al., *Adversarial Filters of Dataset Biases,* in Proceedings of the 37th International Conference on Machine Learning, PMLR 119, 1078-1088 (2020).

informatics project could lead to enough data to unlock the ability to not just evaluate models and verify AI systems, but actually train (or at least fine-tune pre-trained) models on prediction and decision task data derived from legal text.

Data curation should be designed such that the data will be both useful in the near-term for today's AI models, and crucially, so it will also likely be of increasing value as a function of the increase in AI capabilities.[233]

III. CONTRACTS & STANDARDS: *HUMAN-AI* ALIGNMENT

Specifying what we want is hard. The difficulty compounds when we hand inadequate specifications over to powerful optimizers like AI systems that do not share our ontology of abstract concepts.[234] *Law Informs Code* with a tradition of methods (for drafting and interpreting statutes and contracts) that facilitates communicating what a human wants an agent to do.[235]

For most systems, the human deploying the system would like it to obey public laws,[236] but that is not the originating purpose of any practical AI deployment. The purpose is to automatically answer your questions, or to serve as your personal assistant[237] scheduling meetings and booking flights on your behalf,[238] or to drive your car,[239] or to produce beautiful images on command.[240] Something directly useful to you. Contracts can help.[241]

One way of describing the deployment of an AI system is that some 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 we align *AI* with *P*.[242]

Contracts memorialize a shared understanding between parties regarding *value-action-state* tuples. It is not possible to create a complete contingent contract between *AI* and *P* because *AI*'s training process is never comprehensive of every *action-state* pair (that *P* may have a value

---

[233] In particular, in natural language processing on long documents, and through leveraging offline and human-in-the-loop reinforcement learning; *see infra* Section II.C.1.i.
[234] *See, e.g.*, Alex Turner, Neale Ratzlaff & Prasad Tadepalli, *Avoiding Side Effects in Complex Environments*, in Advances in Neural Information Processing Systems 33 21406 (2020); Victoria Krakovna et al., *Avoiding Side Effects by Considering Future Tasks*, in Advances in Neural Information Processing Systems 33 19064 (2020).
[235] *Law also Informs Code* by specifying what AI systems should not do, in order to provide a broader knowledge base of how to reduce externalities and promote coordination and cooperation within a society (*see infra* Section IV).
[236] *See infra* Section IV.
[237] *See, e.g.,* Askell *Laboratory for Alignment*.
[238] *See, e.g.,* Jessy Lin et al., *Inferring Rewards from Language in Context* (2022).
[239] *See, e.g.,* W. Bradley Knox et al., *Reward (Mis)design for Autonomous Driving*, arxiv.org (Mar. 11, 2022), https://arxiv.org/abs/2104.13906.
[240] *See, e.g.,* Aditya Ramesh et al., *Hierarchical Text-Conditional Image Generation with CLIP Latents* (2022).
[241] *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:2208.10469 (2022) (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*, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (2019) [Hereinafter Hadfield-Menell *Incomplete Contracting*].
[242] *See, generally,* Hadfield-Menell *Incomplete Contracting*.

judgment on) that *AI* will see in the wild once deployed.[243] Although it is also practically impossible to create complete contracts between humans, contracts still serve as incredibly useful customizable commitment devices to clarify and advance shared goals. This works because the law has developed mechanisms to facilitate commitment amongst ambiguity. Gaps within contracts – *action-state pairs* without a *value* – are often filled by the invocation of frequently employed standards (e.g., "material" and "reasonable"[244]). These standards could be used as modular (pre-trained model) 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 systems could view contracts (and their creation, implementation, evolution,[245] and enforcement) as (model inductive biases and data) guides to navigating webs of inter-agent[246] obligations.[247] 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*, parties will often not share full information about their preferences because they perceive that information sharing to be strategically disadvantageous *ex ante*: they may plan to attempt to further their goals *ex post*. Dropping the strategic nature of the relationship removes this incentive to withhold useful information.[248] Second, *during the term of the contract*, parties will not be conducting economic analyses of whether breach is more favorable than performance.[249] When we remove the enforcement concerns from the contracts, it removes downfalls such as these,[250] but it does not deprive the *Law Informs Code* approach of the utility of the tools that have evolved to facilitate effective contracting, e.g., extra-contractual standards used to fill "contract" gaps in informing *AI* what to do for *P*.

A. Standards

[243] *See*, Hadfield-Menell *Incomplete Contracting*. In some cases, e.g., 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*, A.I. & L. (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 an action-state space for this approach to be feasible.

[244] Alan D. Miller & Ronen Perry, *The Reasonable Person*, 87 NYU L. Rev. 323 (2012); Karni A. Chagal-Feferkorn, *The Reasonable Algorithm*, U. Ill. JL 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 *Reasonableness*; 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).

[245] *See*, Matthew Jennejohn, Julian Nyarko & Eric Talley, *Contractual Evolution*, 89 U. Chi. L. Rev. 901 (2022).

[246] legal entities, e.g., humans and corporations.

[247] Charles Fried, *Contract as Promise: A Theory of Contractual Obligation* (Harv. Univ. Press, 1981).

[248] *See,* Anthony J. Casey & Anthony Niblett, *Self-Driving Contracts,* in The Journal of Corporation Law (2017) [Hereinafter Casey, *Self-driving*].

[249] *See, e.g.*, Oliver Wendell Holmes, Jr., *The Path of the Law*, 10 HARV. L. REV. 457, 462 (1897) ("[t]he 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 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.

[250] It reduces the importance of legal concepts related to the enforcement of contracts, e.g., mutual mistake and impossibility; *see,* Casey, *Self-driving*. However, doctrines of this nature are still useful to the AI for understanding the context in which existing contracts data was created and their meaning amongst counterparties.

A key engineering design principle, especially for building complicated computational systems, is to leverage modular, reusable abstractions that can be flexibly plugged into a diverse set of systems.[251] In the contracts world, standards are modular, reusable abstractions employed to align agents engaged in inherently incompletely specified relationships and uncertain circumstances.[252] Foundational pre-training of deep learning models, before they are fine-tuned to application-specific tasks, is a potential pathway for embedding the concepts of legal standards, and associated downstream behaviors exhibiting those standards, into AI models. Rules describing discrete logical contractual terms, and straightforward regulatory specifications, can be bolted onto the overall automated system,[253] outside of (the end-to-end differentiable) deep learning model(s), but standards require more nuanced approaches.

For humans, rules are generally more expensive to make but then cheaper to use (because it is clearer whether an action follows a rule), relative to standards, which are more costly than rules to use (because when choosing an action in real-time there is high uncertainty about whether the action is *ex-post* going to comply with the standard).[254] For AI systems, standards are more expensive to instill (and validate) in the system through extensive machine learning training and validation, but then cheaper to deploy because, through models that generalize, they scale to unenumerated state-action pairs. In the *Law Informs Code* use-case, in contrast to their legal creation and evolution,[255] 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 computational process of iteratively defining standards through judicial opinion about their particular case-specific application, and regulatory guidance, can be leveraged as the AI's starting point.

B. An Example: Fiduciary Duty

If law is the applied philosophy of multi-agent alignment, fiduciary law is the branch of that applied philosophy specifically concerned with a principal – a human with less control or information related to the provision of a service – with a fiduciary delegated to provide that service.[256] Fiduciary duties are imposed on powerful agents (e.g., directors of corporations and

---

[251] *See, e.g.*, François Chollet, *Deep Learning with Python, Second Edition* (Manning, 2021); Olivier L. de Weck, Daniel Roos & Christopher L. Magee, *Engineering Systems: Meeting Human Needs in a Complex Technological World* (MIT Press, 2011).
[252] *See infra* Section II. A. 3.
[253] *See, e.g.*, computational logic-based representations of legal institutions and their relationships, *e.g.*, King et al., *A Framework for Governing Institutions*, in Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems, 473-481 (ACM, 2015).
[254] Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 Duke Law Journal 557-629 (1992) at 1 ("Rules typically are more costly than standards to create, whereas standards tend to be more costly for individuals to interpret when deciding how to act and for an adjudicator to apply to past conduct.").
[255] *See, e.g.*, Dale A. Nance, *Rules, Standards, and the Internal Point of View*, 75 FORDHAM L. REV. (2006); Sheppard *Reasonableness*.
[256] 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. 80b; and 17 CFR 275), 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), *see, 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*, (2008), and citations therein, *e.g.*, *see*, Ledbetter v. First State Bank & Trust Co.,

investment advisers[257]) to align their behavior with the wellbeing of the humans for which they are providing services (e.g., corporate shareholders, and investment clients). Fiduciary standards are an empirically and theoretically rich area of law. The concept of fiduciary duty is widely deployed across financial services,[258] business more generally, healthcare, and more. Legislators, regulators, and self-regulatory organizations recognize that it is impossible to create complete contracts between agents (e.g., corporate boards, and investment advisors) and the humans they serve (e.g., shareholders, and investors). AI research also grapples with the impossibility of fully specified *state-action-reward* spaces for training AI agents that generalize to all potentially relevant circumstances.[259] Complete contingent contracts (even if only implicitly complete) between an AI system and the human(s) it serves are implausible for any systems operating in a realistic environment.[260] Fiduciary duties are often seen as part of a solution to the incompleteness of contracts between shareholders and corporate directors, and between investors and their advisors.[261]

---

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 CAL. 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.

[257] Securities laws help align powerful agents (e.g., investment advisors) with their less informed human principals (e.g., investment clients) through fiduciary obligations. As AI becomes more generally capable, securities laws will increasingly apply directly to AI systems because buying, managing, offering, and selling securities are key vectors through which sufficiently advanced automated systems will interact within the broader world. Expanding the purview of the SEC over many AI systems could help enforce human-AI alignment.

[258] In addition to fiduciary duty, there are at least five additional parallels between AI alignment and financial services law. *(1.)* We are attempting to align AI intentions with preferences of groups of humans – "Environmental, Social, Governance" investment products attempt to codify human values and securities regulators find that "greenwashing" is common. *(2.)* We are attempting to manage complex novel AI systems with unpredictable behaviors – financial markets regulators routinely grapple with managing emergent behavior of complex adaptive systems, *see, e.g.,* Yesha Yadav, *The Failure of Liability in Modern Markets*, 102 Va. L. Rev. 1031 (2016). *(3.)* We are witnessing AI power concentrating in private firms with significant data and computing resources, such as large online advertising companies – we have seen the same thing happen over the past few decades in financial markets with the rise of private firms with significant data and computing resources, such as "platform hedge fund" firms managing tens of billions of dollars and the increasing difficulty of launching new alpha-seeking investment firms in that oligopoly-esque environment. *(4.)* Self-regulation is being discussed by AI companies (*see, e.g.* this early effort by a consortium of AI research companies, Cohere, OpenAI, and AI21 Labs: https://openai.com/blog/best-practices-for-deploying-language-models/) – FINRA is an example of a powerful self-regulatory body in financial services. *(5.)* Another lesson from financial regulation: corporate disclosure rules can work well but regulators should fight the urge toward them becoming boilerplate and devolving into performative box-checking.

[259] AI alignment research recognizes a similar problem, *see, e.g.,* Abram Demski & Scott Garrabrant, *Embedded Agency* (2020) at 6, https://arxiv.org/abs/1902.09469 ("the 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, eds. Victor Callaghan, Jim Miller, Roman Yampolskiy & Stuart Armstrong (2017). Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol 1, 501–507 (2019) at 505 ("AI development is not a formal profession. Equivalent fiduciary relationships and complementary governance mechanisms do not exist for private sector AI developers.").

[260] *See*, Hadfield-Menell *Incomplete Contracting.*

[261] Michael C. Jensen & William H. Meckling, *Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure*, Journal of Financial Economics, Vol 3, Issue 4, 305-360 (October 1976); Deborah A. DeMott,

However, fiduciary duty adds value beyond more complete contracts.[262] Even if parties could theoretically create a complete contract up front, there is still something missing: it's not a level playing field between the entities creating the contract (parallel: AI has access to more information and computing power than humans). Contracts generally assume the parties are strategic, negotiating during the contract creation (parallel: the AI objective design) process, but the human-AI relationship is not fundamentally strategic during that design process. Contracts are generally assumed to be created between equals, whereas fiduciary duties are explicitly placed on the party entrusted with more power or knowledge. Fiduciary duty captures this dynamic of asymmetric parties that need guardrails to facilitate alignment of a principal with their agent.

Fiduciary duty goes beyond the explicit contract and helps guide a fiduciary in *a priori* unspecified state-action-value tuples;[263] 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."[264] Contrary to a fiduciary relationship, "no party to a contract has a general obligation to take care of the other, and neither has the right to be taken care of."[265] 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."[266]

A fiduciary duty has two primary components: a duty of loyalty and a duty of care.[267] The duty of care could be interpreted to describe the capability of the AI to accomplish useful behavior

Breach of Fiduciary Duty: On Justifiable Expectations of Loyalty and Their Consequences, 48 Arizona L. Rev. 925-956 (2006); SEC v. Capital Gains Res. Bureau, Inc., 375 U.S. 180, 194-95 (1963); 15 U.S.C. 80b; 17 CFR 275.

[262] Alexander Styhre, *What We Talk About When We Talk About Fiduciary Duties: The Changing Role of a Legal Theory Concept in Corporate Governance Studies*, Management & Organizational History 13:2, 113-139 (2018) [Hereinafter, Styhre, *What We Talk About*]; Arthur B. Laby, *The Fiduciary Obligation as the Adoption of Ends*, 56 Buff. L. Rev. 99 (2008).

[263] Styhre, *What We Talk About.*

[264] *See, e.g.,* D. G. Smith, *Critical Resource Theory of Fiduciary Duty*, 55 Vanderbilt L. Rev. 1399–1497 (2002) at 1410; Deborah DeMott, *Beyond Metaphor: An Analysis of Fiduciary Obligation*, Duke Law Journal (1988) at 882 (The "fiduciary's duties go beyond mere fairness and honesty; they oblige him to act to further the beneficiary's best interests.").

[265] Tamar Frankel, *Fiduciary Law*, 71 California L. Rev. (1983) at 880, https://www.jstor.org/stable/3480303.

[266] *Id.* at 830. According to some legal scholars, fiduciary law has arguably been an important contributor to the economic growth in modern societies, "Exchange 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." Tamar Frankel, *The Rise of Fiduciary Law*, Boston Univ. School of Law, Public Law Research Paper No. 18-18, https://ssrn.com/abstract=3237023 (August 22, 2018) at 11.

[267] *See*, G. Rauterberg & E. Talley, *Contracting Out of the Fiduciary Duty of Loyalty: An Empirical Analysis of Corporate Opportunity Waivers*, Columbia L. Rev. 117 (5): 1075–1151 (2017) (Discussing the distinction between duty of loyalty and duty of care in the context of Delaware corporate law.).

for humans.[268] The 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 becomes more capable and agentic.[269]

1. Information Expression

An example of how legal enforcement expresses information, *in and of itself*,[270] is what an AI can glean from the focus on *ex ante* (human and corporate) deterrence with a default rule for how any gains to actions 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*."[271] Other means of legal deterrence can center more around *post-hoc* financial sanction or incapacitation. If embedded in AI model pre-training processes, standards that pursue deterrence by attempting to remove the ability to share in the gains of negative behavior(s) could guide an AI agent upheld to this standard toward, "*the 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."*[272]

2. A Spectrum

Within financial services, there is a spectrum of fiduciary obligations, e.g., a trustee has significant obligations, while an index provider has a more tenuous fiduciary obligation to the investors in any funds tracking their index.[273] Analogously, fiduciary duty can be a useful standard both for today's AI models and for much more capable models that may be developed over the coming years. Today's deployed AI systems are more similar to the index provider powering a simple rule-based investment strategy, like an exchange-traded fund tracking a standard S&P 500

---

[268] The alignment problem is broken down by some into outer and inner alignment, *see, e.g.*, Evan Hubinger et al., *Risks from learned optimization in advanced machine learning systems* (2019). Eliciting human preferences, distilling them into suitable computational representations, and training an AI agent to understand them solves the "outer alignment" problem by aligning humans and the design of the objectives of their AI. Robustly implementing that AI design specification into behaviors of the AI so that it reliably reflects the human preferences and does not optimize for its own goals not shared by the human solves the "inner alignment" problem by aligning the design of the AI agent with its observed and potential behavior. The AI is fully loyal to the human if we solve inner alignment. In the context of fiduciary standards, outer alignment could potentially be interpreted as the duty of care, and inner alignment as the duty of loyalty.

[269] *See, e.g.*, Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?* (2022) at 4-7 https://arxiv.org/pdf/2206.13353.pdf.

[270] *See infra* Section II. A. 1.

[271] Robert H. Sitkoff, *The Economic Structure of Fiduciary Law*, Boston University L. Rev. (2011) at 1049 [Hereinafter, Sitkoff *The Economic Structure*].

[272] Sitkoff *The Economic Structure* at 1049.

[273] *SEC Requests Information and Comment on Advisers Act Regulatory Status of Index Providers, Model Portfolio Providers, and Pricing Services* (2022).

index,[274] whereas future more advanced AI systems are likely to be more analogous to something like a Trustee administering investments in complicated private equity transactions. We would aim to dial up the fiduciary obligations for more advanced AI systems.

Another way of looking at this: assuming increased capabilities, AI could enable fiduciary duties to be more broadly applied across digital services. In scenarios where an agent is trusted to adopt a principal's objectives, standards that help ensure the agent can be trusted could be foundational to application-specific training processes. In addition to traditionally clear-cut fiduciaries (such as investment advisers), automated personal assistants, programming partners,[275] and other emerging AI-driven services could be designed to exhibit fiduciary obligations toward their human clients.[276] Corresponding advancing capabilities of AI could eventually make this possible by enabling scalability of personalized advice (the basis of the duty of care), while the advancing AI capabilities make the duty of loyalty component increasingly salient.[277]

3. Toward Implementation

One possibility for implementing fiduciary standards is to develop a base level pre-training process for learning the standard across various contexts, while using existing human-AI alignment techniques, such as reinforcement learning from human feedback, as the "contract" component, e.g., by personalizing the AI reward functions to the preferences of the individual human(s) that the AI system is working on behalf of.[278] To learn the standard across various contexts, there are many existing relationships that can be converted to data and training processes, "*Fiduciary principles govern an incredibly wide and diverse set of relationships, from personal relationships and professional service relationships to all manner of interpersonal and institutional commercial relationships. Fiduciary principles structure relationships through which children are raised, incapable adults cared for, sensitive client interests addressed, vast sums of money invested, businesses managed, real and personal property administered, government functions performed, and charitable organizations run. Fiduciary law, more than any other field, undergirds the increasingly complex fabric of relationships of interdependence in and through which people come*

---

[274] SEC, *Commission Interpretation Regarding Standard of Conduct for Investment Advisers* (2019).

[275] *See, e.g.,* Mark Chen et al., *Evaluating Large Language Models Trained on Code*, arXiv:2107.03374 (2021).

[276] It has been argued that the digital age, "has given rise to new fiduciary relationships created by the explosion of the collection and use of personal data. The relationships of trust between end-users and online service providers need not be identical to traditional professional relationships in all respects." Jack M. Balkin, *Information Fiduciaries and the First Amendment,* U.C. Davis. L. Rev. (2016) at 1221. In the context of data privacy, Jack M. Balkin, *The Fiduciary Model of Privacy*, Harvard L. Rev. Forum, Vol. 134, No. 1 (November 2020) argues that technology companies should have fiduciary duties toward their end users, including a duty of confidentiality, a duty of care, and a duty of loyalty.

[277] *See, e.g.,* Michael K. Cohen, Marcus Hutter & Michael A. Osborne, *Advanced Artificial Agents Intervene in the Provision of Reward*, AI Magazine (2022) https://onlinelibrary.wiley.com/doi/full/10.1002/aaai.12064; Richard Ngo, *The Alignment Problem From a Deep Learning Perspective* (Aug 10, 2022) https://www.alignmentforum.org/posts/KbyRPCAsWv5GtfrbG/the-alignment-problem-from-a-deep-learning-perspective.

[278] For the part of the alignment problem under what we could call the duty of care component, a common approach involves eliciting human preferences in a form legible to AI and using reinforcement learning to train the AI to exhibit behaviors that reliably reflect those preferences; *see, e.g.,* Long Ouyang et al. *Training Language Models to Follow Instructions with Human Feedback*, arxiv.org (Mar. 4, 2022), https://arxiv.org/pdf/2203.02155.pdf.

to rely on one another in the pursuit of valued interests."[279] For instance, there is a rich set of fiduciary behavior from corporate directors (who serve as fiduciaries to shareholders) and investment advisers (who serve their clients) from which AI could learn. Corporate officers and investment advisors face the issue of balancing their own interests, the interests of their principals, and the interests of society at large. Unlike most human decision-making, corporations' and investment advisers' behaviors are well documented and are often made by executives with advisors that have deep knowledge of the relevant law. This opens up the possibility of tapping into this observational data to learn best (and worst) practices, and train agents accordingly. [280]

This could involve codifying examples of fiduciary behavior into a structured format to train models. These could include both "gold-standard" human labeled data and automated data structuring sampled and selectively human validated for correctness. The aim would be to use this semi-structured data to design self-supervised learning methods that could be applied across the relevant case law, regulatory guidance, and self-regulatory data in order to train models to learn correct and incorrect fiduciary behavior across many contexts.

One of the first goals should be to develop public benchmark datasets and simulation environments for specific legal standards and policy areas, with the ultimate goal being the widespread adoption of a validation practice demonstrating any given AI system's "understanding" of the legal and regulatory standards relevant to its potential deployment.[281]

IV. PUBLIC LAW: *SOCIETY-AI* ALIGNMENT

If we succeed with the *Law Informs Code* (contracts and standards) approach in increasing the alignment of one AI system to one human (or a small group of humans), we will have a more useful and potentially more *locally* reliable system. However, all else equal, this may decrease the expected global reliability and safety of the system as it interacts with the broader world, e.g., by increasing the risk of the system maximizing the welfare of a small group of powerful people.[282] There are many more objectives and many more people that must be considered.

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.[283] 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.[284] Moving beyond the

[279] Paul B. Miller, *The Identification of Fiduciary Relationships* (February 6, 2018); Evan J. Criddle, Paul B. Miller & Robert H. Sitkoff, eds., *The Oxford Handbook of Fiduciary Law* (Oxford University Press, 2019).
[280] *See infra* Section II. C. 2. ii. for an outline of the methodology for generating the training data from observational legal textual data.
[281] *See infra* Section II. B. 2.
[282] *See, e.g.,* Langdon Winner, *The Whale and the Reactor: A Search for Limits in an Age of High Technology*, University of Chicago Press (2010); Mark Coeckelbergh, *The Political Philosophy of AI* (2022).
[283] Iason Gabriel, *Artificial Intelligence, Values, and Alignment*, 30 MINDS & MACHINES 411 (2020) [Hereinafter Gabriel, *Values*.]; S. Blackburn, *Ruling Passions: An Essay in Practical Reasoning* (Oxford University Press, 2001).
[284] Gabriel, *Values*.

problem of aligning AI with a single human, aligning advanced AI with society is considerably more difficult,[285] but it is necessary as AI deployment has broad effects.[286]

Eliciting and synthesizing human values systematically is an unsolved problem that philosophers and economists have labored on for millennia.[287] Even if we miraculously "solved" that – and AI capabilities research made further progress in multi-agent contexts[288] – much of the technical AI alignment research is still focused on the solipsistic "single-single" problem of single human and a single AI.[289] The pluralistic dilemmas stemming from "single-multi" (a single human and multiple AIs) and especially "multi-single" (multiple humans and a single AI[290]) and "multi-multi" situations are critical.[291] When we are attempting to align multiple humans with one or more AI system, we need overlapping endorsements of AI behaviors,[292] but there is no consensus social choice mechanism to aggregate preferences and values across humans[293] or time.[294] 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. [...] Each set of decisions poses difficult ethical dilemmas with major consequences for AI behavior, with some decision options yielding pathological or even catastrophic results."[295]

---

[285] *See, e.g.*, Andrew Critch & David Krueger, *AI Research Considerations for Human Existential Safety (ARCHES)* (2020) [Hereinafter, Critch, *AI Research Considerations*]; Eliezer Yudkowsky, *Coherent Extrapolated Volition*, Singularity Institute for Artificial Intelligence (2004); Hans De Bruijn & Paulien M. Herder, *System and Actor Perspectives on Sociotechnical Systems*, IEEE Transactions on systems, man, and cybernetics-part A: Systems and Humans 39.5 981 (2009); Jiaying Shen, Raphen Becker & Victor Lesser, *Agent Interaction in Distributed POMDPs and its Implications on Complexity*, in Proceedings of the Fifth International Joint Conference on Autonomous Agents and Multiagent Systems, 529-536 (2006).

[286] *See*, Ben Wagner, *Accountability by Design in Technology Research*, Computer Law & Security Review, 37 105398 (2020); Roel Dobbe, Thomas Krendl Gilbert & Yonatan Mintz, *Hard Choices in Artificial Intelligence*, Artificial Intelligence, 300 103555 (2021).

[287] *See, e.g.*, Gabriel, *Values;* Ariela Tubert, *Ethical Machines*, 41 Seattle U L Rev 1163 (2017); Amartya Sen, *Rationality and Social Choice*, The American Economic Review 85.1 (1995).

[288] *See, e.g.*, Max Jaderberg et al., *Human-level Performance in 3D Multiplayer Games with Population-based Reinforcement Learning*, Science 364.6443 859-865 (2019); Hengyuan Hu, Adam Lerer, Alex Peysakhovich & Jakob Foerster, *"Other-Play" for Zero-Shot Coordination*, In International Conference on Machine Learning, 4399-4410 (2020); Johannes Treutlein, Michael Dennis, Caspar Oesterheld & Jakob Foerster, *A New Formalism, Method and Open Issues for Zero-shot Coordination*, In International Conference on Machine Learning, 10413-10423 (2021); Phillip Christoffersen, Andreas A. Haupt & Dylan Hadfield-Menell, *Get It in Writing: Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL*, arXiv:2208.10469 (2022); Chongjie Zhang & Julie A. Shah, *Fairness in Multi-agent Sequential Decision-making*, in Advances in Neural Information Processing Systems 27 (2014); Siqi Liu et al., *From Motor Control to Team Play in Simulated Humanoid Football*, SCIENCE ROBOTICS Vol 7, Issue 69 (Aug 31 2022) (They demonstrate agents learning coordination in a relatively complex multi-agent environment.).

[289] *See*, Critch, *AI Research Considerations*.

[290] *See, e.g.*, Arnaud Fickinger et al., *Multi-Principal Assistance Games: Definition and Collegial Mechanisms*, arXiv:2012.14536 (2020).

[291] Critch, *AI Research Considerations*.

[292] *See, e.g.*, Gabriel, *Values.*

[293] *See*, Amartya Sen, *Collective Choice and Social Welfare*, Harv. Univ. Press (2018); G. Arrhenius, *An Impossibility Theorem for Welfarist Axiologies*, Economics & Philosophy, 16(2), 247–266 (2000); Seth D. Baum, *Social Choice Ethics in Artificial Intelligence,* AI & SOCIETY 35.1, 165-176 (2020); Critch, *AI Research Considerations*; Gabriel, *Values*.

[294] Tyler Cowen & Derek Parfit, *Against the Social Discount Rate*, in Peter Laslett & James S. Fishkin (eds.) Justice Between Age Groups and Generations, Yale University Press 144 (1992).

[295] Seth D. Baum, *Social Choice Ethics in Artificial Intelligence,* AI & SOCIETY 35.1, 165-176 (2020) at 1.

Rather than attempting to reinvent the wheel in ivory towers or corporate bubbles, we should be inspired by democratic mechanisms and resulting law.[296]

In addition to *Law Informing Code* through standards and through interpretation methods that facilitate specifying what a human wants an agent to do,[297] *Law Informs Code* with a constantly updated and verified knowledge base of societally endorsed human preferences on what AI systems should not do, in order to reduce externalities and promote coordination and cooperation.[298]

A. AI Ethics and Moral Machines

The *Law Informs Code* approach should be the core alignment framework, with efforts to embed (ever-contested) "ethics" into AI as a key complement.[299] Human intuition, our "common sense morality," often falters in situations that involve decisions about groups unlike ourselves.[300] Due to this "Tragedy of Common-Sense Morality,"[301] instead of attempting to replicate common sense morality, scholars have suggested various academic philosophical theories to address the AI-society alignment dilemma.[302]

---

[296] However, 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*, Proceedings of the National Academy of Sciences 112.33 (2015); Mark Coeckelbergh, *The Political Philosophy of AI* (2022) at 62-92; Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, Public Affairs (2019).

[297] *See infra* Section III.

[298] Although this Article is not focused on how *Law Governs AI*, the *Law Informs AI* agenda suggests a novel policy approach (as a thought experiment) to governing AI's relationships to humans and the physical world: wrapping all agentic AI systems in their own legal entities, e.g., a Corporation or a Limited Liability Company. A public policy, then, to further align AI with humans would be to enforce that the legal entity has verified human shareholders. The corporation is, to a large extent, a mechanism designed to reduce the principal-agent problem between shareholders and managers (DGCL §141(a) ("The business and affairs of every corporation organized under this chapter shall be managed by or under the direction of a board of directors...."), so with humans as the shareholders the corporate form could help align the corporate AI "management." Regardless of whether wrapping the system in a legal entity would be helpful, under current law, sufficiently advanced AI systems would be able to utilize legal business entities as the key vector through which they conduct their affairs, e.g., to employ humans, to sue other entities, to purchase goods (*see, e.g.*, Shawn Bayern, *Are Autonomous Entities Possible*, NW. U. L. REV. Online 114 23 (2019); Lynn M. LoPucki, *Algorithmic Entities*, Vol 95, Issue 4, 887-953 Washington University L. Rev. (2018); Shawn Bayern, *The Implications of Modern Business Entity Law for the Regulation of Autonomous Systems*, 19 STAN. TECH. L. REV. 93, 104 n.43 (2015); Shawn Bayern, *Of Bitcoins, Independently Wealthy Software, and the Zero-Member LLC*, 108 NW. U. L. REV. 1485, 1496–97 (2014)). Reducing the risk that this potentially inevitable state of affairs leads to bad outcomes for humans may imply the same conclusion: we should strictly enforce that business entities have human shareholders.

[299] *See, e.g.*, Joshua Walker, *Is 'Ethical AI' a Red Herring?* 36 Santa Clara High Tech LJ 445 (2019).

[300] Joshua Greene, *Moral Tribes: Emotion, Reason, and the Gap Between US and Them*, Penguin Press (2013) [Hereinafter Greene, *Moral Tribes*].

[301] Greene, *Moral Tribes*.

[302] *See, e.g.*, Wendell Wallach & Colin Allen, *Moral Machines: Teaching Robots Right From Wrong* (2009); Michael Anderson & Susan L. Anderson, *Machine Ethic*s, Cambridge University Press (2011); Edmond Awad et al., *Computational Ethics*, In Trends in Cognitive Sciences (2022); James H. Moor, *Just Consequentialism and Computing*, Ethics and Information Technology 1, 61–65 (1999); Heather M. Roff, *Expected Utilitarianism* (2020); Elizabeth Gibney, *The Battle for Ethical AI at the World's Biggest Machine-learning Conference*, Nature (2020); Dan Hendrycks et al., *Aligning AI With Shared Human Values* (2021); National Academies of Sciences, Engineering, and

We provide six reasons why ethics is likely not the best primary framework for AI alignment. *First*, most ethical theory is too imprecise to be practically useful for building AI systems;[303] it does not meet our first desired characteristic of an alignment framework.[304]

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, "The 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."[305] Two corollaries to these first two issues are that we cannot validate the ethics of an AI system in any widely agreed-upon manner,[306] and there is no database of empirical applications (especially not one with sufficient ecological validity[307]) that can be leveraged by machine learning processes.[308] Law is validated in a widely agreed-upon manner and has massive databases of empirical application with sufficient ecological validity.

Third, ethics, probably by its nature, lacks settled precedent across, and even within, theories.[309] 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."[310]

Medicine, *Fostering Responsible Computing Research: Foundations and Practices*, Washington, DC: The National Academies Press (2022); Joshua Greene et al., *Embedding Ethical Principles in Collective Decision Support Systems*, Thirtieth AAAI Conference on Artificial Intelligence (2016).

[303] *See, e.g.*, Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol 1, 501–507 (2019) 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"); R. Clarke, *Principles and Business Processes for Responsible AI*, Computer Law and Security Review (2019); Jessica Morley et al., *Ethics as a Service: A Pragmatic Operationalisation of AI Ethics*, 31.2 Minds and Machines 239-256 (2021); J. van den Hoven, *Computer Ethics and Moral Methodology*, Metaphilosophy 28, 234–248 (1997); W. B. Gallie, *Essentially Contested Concepts*, Proc. Aristot. Soc. 56, 167–198 (1955); H. S. Richardson, *Specifying Norms as a way to Resolve Concrete Ethical Problems*, Philos. Public Aff. 19, 279–310 (1990).

[304] *See infra* Section I for the framework requirements.

[305] Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol 1, 501–507 (2019) at 503; K. Shilton, *Values Levers: Building Ethics Into Design*. Sci. Technol. Hum. Values 38, 374–397 (2013).

[306] *See, e.g.*, Anne Gerdes & Peter Øhrstrøm, *Issues in robot ethics seen through the lens of a moral Turing test*, Journal of Information, Communication and Ethics in Society (2015); J. Van den Bergh & D. Deschoolmeester, *Ethical Decision Making in ICT: Discussing the Impact of an Ethical Code of Conduct*, Commun. IBIMA, 127497 (2010); B. Friedman, D. G. Hendry & A. Borning, *A Survey of Value Sensitive Design Methods*, Found. Trends Hum. Comp. Interact. 11, 63–125 (2017); Brent Mittelstadt, *Principles alone cannot guarantee ethical AI*, Nature Machine Intelligence, Vol 1, 501–507 (2019).

[307] *See, e.g.*, Martin T. Orne & Charles H. Holland, *On the Ecological Validity of Laboratory Deceptions*, International Journal of Psychiatry 6, no. 4, 282-293 (1968).

[308] *See infra* Section II. C. 2.

[309] *See, e.g.*, Gabriel, *Values* at 425 ("it 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.").

[310] Miles Brundage, *Limitations and risks of machine ethics*, Journal of Experimental and Theoretical Artificial Intelligence, 26.3, 355–372 (2014) at 369.

Fourth, even if AI designers (impossibly) agreed on one ethical theory (or ensemble of underlying theories[311]) being "correct," there is no mechanism to align humans around that theory.[312] In contrast, in democracies, law has legitimate authority imposed by widely accepted institutions,[313] and serves as a coordinating focal point of values to facilitate human progress.[314]

Fifth, even if AI designers (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 (usually, improving) ethical standards; 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 on a positive trajectory. Therefore, we do not want to lock in today's ethics without a clear and trustworthy update mechanism.[315] In stark contrast, law is constantly formally revised to reflect the evolving will of citizens.[316] If AI is designed to use law as a key source of alignment insight, this would build in an automatic syncing with the latest iteration of synthesized and validated societal value preference aggregation.[317]

Sixth, veering into the intersection of *Law Informs Code* and *Law Governs Code*, there is a practical reason law may be better suited as an alignment framework than ethics. Bryan Casey points out that most of the entities deploying impactful AI systems are companies; companies are profit-maximizers; and profit-maximizers, when faced with how to add constraints to their core value proposition, are liability-minimizers rather than morality-maximizers.[318] Casey concludes, optimistically, that, "We, 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."[319] If law informs powerful AI, engaging

[311] *See, e.g.*, Toby Newberry & Toby Ord, *The Parliamentary Approach to Moral Uncertainty*, Technical Report # 2021-2, (Future of Humanity Institute, University of Oxford, July 15, 2021) https://www.fhi.ox.ac.uk/wpcontent/uploads/2021/06/Parliamentary-Approach-to-Moral-Uncertainty.pdf; William MacAskill, *Practical ethics given moral uncertainty*, Utilitas 31.3 231 (2019).
[312] *See, e.g.*, John Rawls, *The law of peoples: with, the idea of public reason revisited*, Harv. Univ. Press (1999) at 11-16; Gabriel, *Values* at 425.
[313] *See, generally*, David Estlund, *Democratic authority: A philosophical framework*, Princeton University Press (2009); Gabriel, *Values* at 432.
[314] "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*, Harv. Univ. Press (2017) at 260 [Hereinafter McAdams, *The Expressive Powers of Law*].
[315] *See, e.g.*, William McAskill, *Are we living at the hinge of history?* Global Priorities Institute Working Paper 12 (2020); William McAskill, *What We Owe the Future* (2022).
[316] In fact, modeling the evolution of an area of law (e.g., the "legislative history" of the drafting and enactment of legislation, and subsequent amendments to the statue) as a sequential decision-making process could be a useful method for AI to learn implicit "reward functions" of the citizenry regarding policy areas. For an evolutionary perspective on reward functions, *see, e.g.*, Satinder Singh, Richard L. Lewis & Andrew G. Barto, *Where Do Rewards Come From*, In Proceedings of the annual conference of the cognitive science society, 2601-2606, Cognitive Science Society (2009).
[317] "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." James C. Scott, *Seeing Like a State* (1998) at 357.
[318] *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*].
[319] Casey, *Amoral Machines*, at 1365. *See, also*, Ryan Calo, *Artificial Intelligence and the Carousel of Soft Law*, IEEE Transactions on Technology and Society 2, no. 4 171-174 (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

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 large companies dictate their ethics by fiat.

Ethics should be much more center stage than it currently is for AI researcher training, AI system development best-practices, and AI deployment protocols. Ethics should be part of the technical AI alignment toolbox, and, as we discuss below, it is critical for guiding data selection and processing in legal informatics. At the same time, we agree with John Rawls that, "In a constitutional democracy the public conception of justice should be, as far as possible, independent of controversial philosophical and religious doctrines," and, "the public conception of justice is to be political, not metaphysical."[320] The question for *Law Informing Code*, then, is how to leverage legal data for these purposes.

B. Toward Implementation

Case law teaches AI how to map from democratically determined directives (statutes) to specific implementation, whereas statutes are more useful for embedding world knowledge and human value expressions into AI. Legislation expresses a significant amount of information about the values of citizens,[321] "for example, by banning employment discrimination against LGBT workers, the legislature may communicate pervasive attitudes against such employment practices."[322] And, "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."[323]

Although special interest groups can influence the legislative process, legislation is largely reflective of *citizens' beliefs* because "legislators gain by enacting legislation corresponding to actual attitudes (and actual future votes)."[324] The second-best source of citizens' attitudes is arguably polls, 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 to the sampling techniques. Whereas, 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

---

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.").

[320] John Rawls, *Justice as Fairness: Political Not Metaphysical*, Phil & Pub Aff, at 223, 224–225 (1985); Gabriel, *Values*.

[321] *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. Cm. L. Rev. 1, 66-71 (1995); Cass R. Sunstein, *On the Expressive Function of Law,* Univ of Penn L. Rev., 144.5 (1996); Dhammika Dharmapala & Richard H. McAdams, *The Condorcet Jury Theorem and the Expressive Function of Law: A Theory of Informative Law*, American Law and Economics Review 5.1 1 (2003).

[322] McAdams, *The Expressive Powers of Law* at 137.

[323] Cass R. Sunstein, *On the Expressive Function of Law,* Univ of Penn L. Rev., 144.5 (1996) at 2024.

[324] McAdams, *The Expressive Powers of Law*, at 149.

legislators relied, weighted according to their expert opinion of each poll's reliability."[325] More recent legislation could be interpreted as providing fresher pulse checks on citizen attitudes; 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 tradeoff views* 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."[326] All activities have some level of risk, and making tradeoffs about which activities are deemed to be riskier is ultimately a sociological process with no objectively correct risk ranking.[327] The cultural process of prioritizing risks is reflected in legislation and its subsequent implementation in regulation crafted by domain experts. Finally, 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.[328]

Acknowledging that data contains economic, racial,[329] and gender biases,[330] we frame the challenge of estimating the expressive power of public law[331] broadly to factor in whether views of historically marginalized populations are expressed.[332] Work on fairness, accountability, and transparency of AI[333] should inform research on methods for estimation of a more comprehensive notion of the expressiveness of legal data.[334] Methods are being developed that attempt to improve

[325] McAdams, *The Expressive Powers of Law*, at 146.
[326] McAdams, *The Expressive Powers of Law*, at 138.
[327] *See, e.g.*, on long-term existential risk, Carla Zoe Cremer & Luke Kemp, *Democratising Risk: In Search of a Methodology to Study Existential Risk* (2021).
[328] Richard H. McAdams & Janice Nadler, *Coordinating in the Shadow of the Law: Two Contextualized Tests of the Focal Point Theory of Legal Compliance*, Law & Society Review 42.4 865-898 (2008); Richard H. McAdams, *A Focal Point Theory of Expressive Law*, Virginia Law Review 1649-1729 (2000); Dylan Hadfield-Menell, McKane Andrus & Gillian Hadfield, *Legible Normativity for AI alignment: The Value of Silly Rules*, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 115-121 (2019).
[329] *See, e.g.*, Rashida Richardson, Jason M. Schultz & Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, NYU L Rev. Online 94, 15 (2019); Z. Obermeyer, B. Powers, C. Vogeli & S. Mullainathan, *Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations*, Science, 366 (6464) 447-453 (2019).
[330] *See, e.g.*, Caroline Criado Perez, *Invisible Women: Data Bias in a World Designed for Men* (2019); Joy Buolamwini & Timnit Gebru, *Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*, in Conference on Fairness, Accountability, and Transparency, Proceedings of Machine Learning Research 81: 1–15 (2018).
[331] And across all types of legal data used as part of the legal informatics efforts.
[332] For legal discussions, *see, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Deborah Hellman, *Measuring Algorithmic Fairness*, 106 VA. L. REV. 811 (2020).
[333] *See, e.g.*, Timnit Gebru et al., *Datasheets for Datasets*, COMMC'NS ACM (Dec. 2021); Emily M. Bender & Batya Friedman, *Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science*, 6 TRANSACTIONS ASS'N FOR COMPUTATIONAL LINGUISTICS 587 (2018); Margaret Mitchell et al., *Model Cards for Model Reporting*, in 2019 PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY 220 (2019); William Cai et al., *Adaptive Sampling Strategies to Construct Equitable Training Datasets*, in 2022 PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, 1467–1478 (2022); Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock & Vivek K. Singh, *ABCinML: Anticipatory Bias Correction in Machine Learning Applications*, in 2022 PROC. CONF. ON FAIRNESS, ACCOUNTABILITY, & TRANSPARENCY, 1552–1560 (2022).
[334] *See, e.g.*, McKane Andrus et al., *AI Development for the Public Interest: From Abstraction Traps to Sociotechnical Risks* (2021).

the fairness of machine learning[335] through data preprocessing,[336] adjusting model parameters during training,[337] and adjusting predictions from models that have already been trained.[338] Another issue is that legal data can contain political biases in places where it is purported to be produced by processes fully committed to judicial[339] and agency[340] independence.

Imbuing AI with the capability to understand new statutes is a significant technical challenge.[341] As the state-of-the-art for AI advances, the *Law Informs Code* approach aims to build AI systems that can demonstrate correspondingly advanced legal and regulatory reasoning and statutory interpretation abilities.[342]

V. CONCLUSION

Novel AI capabilities continue to emerge,[343] increasing the urgency to align AI with humans. Law, as the applied philosophy of multi-agent alignment, uniquely fulfills our requirements for an alignment framework.

Alignment is a problem because we cannot directly specify 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. Methodologies for making and interpreting law, for advancing shared goals in new circumstances, have been theoretically refined over centuries. As the state-of-the-art for AI advances, we can set higher bars of demonstrated legal understanding capabilities; if a developer claims their system has advanced capabilities on

---

[335] *See, e.g.,* Reva Schwartz et al, NIST Special Publication 1270, *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*, Natl. Inst. Stand. Technol. Spec. Publ. 1270, 1 - 86 (March 2022) https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf; Michael Kearns & Aaron Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford University Press, 2019) at 57-93.

[336] *See, e.g.*, F. Calmon et al., *Optimized Pre-processing for Discrimination Prevention*, in Advances in Neural Information Processing Systems, 3992–4001 (2017).

[337] *See, e.g.*, M. B. Zafar, I. Valera, M. Gomez-Rodriguez & K. P. Gummadi, *Fairness Constraints: A Flexible Approach for Fair Classification*, in J. Mach. Learn. Res., vol. 20, no. 75, 1–42 (2019).

[338] *See, e.g.*, M. Hardt, E. Price & N. Srebro, *Equality of Opportunity in Supervised Learning*, in Advances in neural information processing systems 3315–3323 (2016).

[339] *See, e.g.,* Neal Devins & Allison Orr Larsen, *Weaponizing En Banc*, NYU L. Rev. 96 (2021) at 1373 ("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, Michael A. Livermore & Daniel N. Rockmore, *The Problem of Data Bias in the Pool of Published US Appellate Court Opinions*, Journal of Empirical Legal Studies 17.2, 224-261 (2020).

[340] 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*, Regulation & Governance 15.3 653-685 (2021).

[341] *See, e.g.,* Nils Holzenberger, Andrew Blair-Stanek & Benjamin Van Durme, *A Dataset for Statutory Reasoning in Tax Law Entailment and Question Answering* (2020). "Ambition must be made to counteract ambition." James Madison, *The Structure of the Government Must Furnish the Proper Checks and Balances Between the Different Departments*, The Federalist Papers: No. 51 (February 8, 1788).

[342] *See infra* Section II B.

[343] *See, e.g.,* Ganguli et al., *Predictability and Surprise in Large Generative Models* (2022) https://arxiv.org/abs/2202.07785; Wenlong Huang et al., *Inner Monologue: Embodied Reasoning through Planning with Language Models*, arXiv:2207.05608 (2022) at 8.

tasks, they should demonstrate correspondingly advanced legal comprehension of the AI. The practices of making, interpreting, and enforcing law have been battle tested through millions of contracts and legal and regulatory actions that have been memorialized in digital format, providing large data sets of examples and explanations, and millions of well-trained active lawyers from which to elicit machine learning model feedback to embed an ever-evolving comprehension of human goals. However, much work, both theoretical and empirical, needs to be undertaken.

For instance, from the theory side, 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.[344] Legal theorizing on these questions would help guide additional legal informatics research, and AI ethics, in particular, is uniquely suited to help address this. We need to conduct a significant amount of AI ethics 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).

This Article developed the ways in which U.S. *law informs code*, but we need to extend this to scale the approach globally.[345]

Mainstream AI capabilities research – while unlocking better AI performance on existing tasks and enabling more powerful and general agents – is likely to have positive externalities on legal informatics alignment performance by enabling greater AI legal reasoning capabilities and legal interpretation skills. However, an area of research with little focus relative to the high value it would deliver to legal informatics capabilities is natural language processing (NLP) of long documents.[346] The current focus of most NLP work is on documents much shorter than the average legal text.[347] To deploy Foundation Models more successfully to legal text data, the AI research community needs to tackle the ability for models to better process and comprehend longer documents. There is promising work in this direction,[348] but it is worthy of more attention and requires more ambition on the length of documents.

We should take a portfolio approach to AI alignment research by experimenting with a variety of empirical and theoretical methodologies. With a concerted effort to advance research at the intersection of AI and law, legal informatics can play a unique role in theoretically framing, and technically implementing, improvements to AI. Legal informatics and AI capabilities research

[344] This is similar to 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*, in Advances in Neural Information Processing Systems 33 19064 (2020).

[345] *See*, Gabriel, *Values* at 426-429; *see, e.g.*, David D. Friedman, *Legal Systems Very Different from Ours* (2012).

[346] *See, e.g.*, Anil et al., *Exploring Length Generalization in Large Language Models* (2022) https://arxiv.org/abs/2207.04901 ("naively finetuning transformers on length generalization tasks shows significant generalization deficiencies independent of model scale").

[347] Congressional bills are routinely hundreds of pages long (*see, generally*, https://www.congress.gov/; John Nay, *Predicting and Understanding Law-making with Word Vectors and an Ensemble Model*, 12 PLOS ONE 1 (2017)).

[348] *See e.g.*, Yunyang Xiong et al., *Nyströmformer: A Nyström-Based Algorithm for Approximating Self-Attention* (2021); Yann Dubois, Gautier Dagan, Dieuwke Hupkes & Elia Bruni, *Location Attention for Extrapolation to Longer Sequences* (2019); Anil et al., *Exploring Length Generalization in Large Language Models* (2022); Iz Beltagy, Matthew E. Peters & Arman Cohan, *Longformer: The Long-Document Transformer* (2020); Nikita Kitaev, Łukasz Kaiser & Anselm Levskaya, *Reformer: The Efficient Transformer* (2020); Jason Phang, Yao Zhao & Peter J. Liu, *Investigating Efficiently Extending Transformers for Long Input Summarization* (2022); Hai Pham et al., *Understanding Long Documents with Different Position-Aware Attentions* (2022).

have the opportunity to inform each other as we iterate toward embedding law, the language of alignment, into AI.