Hardwiring Hercules?

Courtney M. Cox*

It is time to reorient the debate over the right to a human decision. Within that debate, one of the strongest arguments in favor of human decisionmakers are Arguments from Explanation: machines should not replace human decision-makers because AI technology is increasingly and necessarily opaque in a way that prevents machines from giving the sort of explanation required. Or so the humanist argument goes. Meanwhile, machinists argue that most humanist principles have been deflated by the Better Decision Argument, which reframes such principles as grounding not a right to a human decision, but merely to a "better" decision—whether by human or machine.

This Article turns that debate on its head. First, it offers a reason to doubt Arguments from Explanation: human judges sometimes don't know what they ought to do. They have what is called "normative uncertainty." But if they respond to that uncertainty rationally, they will not generally

---

* © 2026 Courtney M. Cox, Associate Professor of Law, Fordham University School of Law. B.A. in Engineering Sciences (Electrical), Yale University; B.Phil., D.Phil. (Philosophy), University of Oxford; J.D., University of Chicago Law School. For helpful comments and conversations, I thank Shyam Balganesh, Doni Bloomfield, Bob Brauneis, Bennett Capers, Mala Chatterjee, Nestor Davidson, Colin Doyle, Janet Freilich, Brian Frye, Katrina Geddes, Michael Goodyear, Abner Greene, James Grimmelmann, Jeremy Hanson, Laura Heymann, Alexander Houghton, Aziz Huq, Martin Kelly, Aniket Kesari, Gowri Krishna, Ela Leshem, Hillel Nadler, Nicholson Price, Richard Re, Lawrence Sager, Sepehr Shahshahani, Chinny Sharma, Jeremy Sheff, Murray Tipping, Nina Varsava, Felix Wu, Benjamin Zipursky, anonymous reviewers, and participants at IPSC, CS+Law, JIPSA, University of Stirling Workshop on Legal Epistemology, Metropolitan Junior Scholars' Workshop, and UConn Law School Faculty Workshop. For discussions on earlier versions and related AI projects, I also thank Tinu Adediran, Amin Ebrahimi Afrouzi, Jeanne Fromer, Mark Lemley, Tejas Narechania, Paul Ohm, Randal Picker, Blake Reid, Pamela Samuelson, Neel Sukhatme, Ari Waldman, Ryan Whalen; and participants at CS+Law, ICAIL, NYC IP & Philosophy Workshop, Suffolk IP & Innovation Workshop, and WIPIP. For research assistance, I thank the Fordham Law Librarians, especially Jamie Taylor, and my incredible research assistants, especially Isabella Ingrao, Ilana Millner, Sydney Crute, Sol Murgui Orsucci, Jasmine Boyer, Kate Bundy, Kevin Burns, Ross Delhagen, Eric Hechler, Nick Manzella, and Lyvia Yan; and for their care and attention, the editors at the *Arizona State Law Journal*, especially Crispin South and Stephen Pearson. I am especially grateful to the organizers for featuring the initial Automating the Uncertain Judge line of work as a closing plenary at IPSC, and to the Fordham Faculty Grant for supporting research that culminated in this project.

provide the kind of explanation demanded of machines. Thus, Arguments from Explanation do not count in favor of human judges, the quintessential example of decisions where an explanation is owed.

But while normative uncertainty gives machinists one advantage, it also undermines the machinists' Better Decision Argument. Machines may offer the illusion of an idealized AI decisionmaker like Dworkin's Hercules. But building a machine usually requires hardwiring objectives. Such hardwiring can preclude rational consideration of normative uncertainty, making it harder for machines to aim at a "better" decision. As a result, normative uncertainty also deflates the machinists' Better Decision Argument.

The best arguments from both camps having been thus undermined, a new question emerges, sharpening the concerns at the heart of the debate. We gain language to diagnose the lingering worry—the pit of dread in one's stomach—that persists in the face of Better Decision Arguments. And we enable lawmakers and regulators to better understand the problem that engineers face but struggle to articulate. As engineers also suspect, hardwiring Hercules is dangerous if you are uncertain what he should do.

INTRODUCTION

The prospect of automating human decisionmakers sometimes inspires dread. But disagreement remains about the source of discomfort—whether it is real or illusory, knee-jerk or justified. Humanist objectors often ground their objections in concerns like transparency and explainability, dignity and autonomy, bias and institutional effects.[1] Machine proponents often counter that the human mind is a black box, that humans can fail to respect dignity, that bias is a problem for humans too—that, in short, all the objectors have shown is a need for a better decision, be that human or machine.[2] The relevant limits are technical, not normative, or so machine proponents claim.[3]

This Article offers something to both sides, and so seeks to reorient the debate. To machine proponents ("machinists"), it offers a new reason to be skeptical of Arguments from Explanation in favor of humans. And to humanist objectors ("humanists"), it offers a new reason to be skeptical of Better Decision Arguments in favor of machines. Flipping the script in this

---

1. *See, e.g.*, Frank A. Pasquale, *Toward a Fourth Law of Robotics: Preserving Attribution, Responsibility, and Explainability in an Algorithmic Society*, 78 OHIO ST. L.J. 1243, 1252–53 (2017) (arguing algorithmic accountability requires a certain kind of explainability and transparency "to ensure that robots and algorithmic agents are traceable to and identified with their creators"); Meg Leta Jones, *The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood*, 47 SOC. STUD. SCI. 216, 231–32 (2017) ("[T]o treat a human in a wholly computational manner reduces the individual's dignity."); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1292–98 (2008) (arguing algorithmic decision-making often runs roughshod over due process and rulemaking values).

2. *See, e.g.*, Bartosz Brożek et al., *The Black Box Problem Revisited: Real and Imaginary Challenges for Automated Legal Decision Making*, 32 A.I. & L. 427, 430 (2024); Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611 (2020) [hereinafter Huq, *Human Decision*]; Cary Coglianese, *A Right to a Better Decision*, REGUL. REV. (June 3, 2024), https://www.theregreview.org/2024/06/03/coglianese-a-right-to-a-better-decision/.

Here and throughout, I use the term "machine" to contrast with "human." I do not use "artificial intelligence" or "AI" because I want to remain technology agnostic except as otherwise stated—that is, I want to avoid making assumptions about what is "under the hood" that use of "AI" is sometimes taken to imply. *See* Margot E. Kaminski & Jennifer M. Urban, *The Right to Contest AI*, 121 COLUM. L. REV. 1957, 1959 n.1 (2021) (noting ambiguity in use of "AI"); Huq, *Human Decision*, *supra*, at 614 (using "machine"). I remain technology agnostic because my analysis focuses on an overlooked normative detail about many different types of machine decisionmakers, not because the technical details aren't important (they are). *Cf.* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 U.C. DAVIS L. REV. 653 (2017); *see also infra* Part III.

3. *See* Huq, *Human Decision*, *supra* note 2, at 651–52; Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 ADMIN. L. REV. 1, 10–13 (2019).

way will sharpen the questions going forward, both technical and normative.

The new reasons to doubt old arguments—on both sides—begin with uncertainty of a particular flavor: namely, normative uncertainty. Normative uncertainty is uncertainty about what one ought to do.[4] To illustrate, consider the trolley problem: Caught on a run-away trolley, should you stay on course and crush five people or turn and kill the one?

I know, cue the groans. You've heard the trolley problem before and yes, self-driving cars encounter it.[5] But I have my reasons for deploying it, starting with its familiarity and ending with the critiques.[6]

In any event, self-driving cars encounter real-life trolley problems, and humans—engineers, regulators, and in some cases, end users—will need to decide what self-driving cars should do when that happens.[7] Set aside the difficulty of who should get to make the decision and suppose that it were up to one of us. The trouble is that there are many reasons we might be unsure what we want the machine to do in such a case—what rule we want it to follow, what criteria to consider—and one of them is normative uncertainty.

You might think of normative uncertainty as that part of our uncertainty that the machine cannot resolve.[8] Machines might be able to help with regular types of nonnormative uncertainty. For example, one reason we might be unsure is that we don't know what the consequences will be: how

---

4. *See generally* WILLIAM MACASKILL, KRISTER BYKVIST, & TOBY ORD, MORAL UNCERTAINTY (2020); TED LOCKHART, MORAL UNCERTAINTY AND ITS CONSEQUENCES (2000) (discussing normative uncertainty in moral decision-making); Courtney M. Cox, *The Uncertain Judge*, 90 U. CHI. L. REV. 739, 739 (2023) [hereinafter Cox, *The Uncertain Judge*] (discussing the "problem of 'normative uncertainty'" in the context of judicial decision-making).

5. Or maybe you've seen the meme. In any event, it began with Philippa Foot, *The Problem of Abortion and the Doctrine of Double Effect*, 5 OXFORD REV. 5 (1967) (introducing the problem); *see also* Judith Jarvis Thomson, *Killing, Letting Die, and the Trolley Problem*, 59 MONIST 204, 206–10 (1976) (canonical treatment of Foot's "trolley problem"). For a sampling of the vast literature, see, for example, FRANCIS M. KAMM, THE TROLLEY PROBLEM MYSTERIES 113 (Eric Rakowski, 1st ed. 2016).

6. Not to mention a few anecdotes that I hope you'll find amusing, like the time an esteemed constitutional law scholar tried to kill me with one.

7. For a sampling of the vast literature, see, for example, Francis Kamm, *The Use and Abuse of the Trolley Problem*, *in* THE ETHICS OF ARTIFICIAL INTELLIGENCE 79, 89 (S. Matthew Liao ed., 2020); Hao Zhan & Dan Wan, *Ethical Considerations of the Trolley Problem in Autonomous Driving: A Philosophical and Technological Analysis*, 15 WORLD ELEC. VEHICLE J. 1 (2024).

8. *Cf.* Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1921 (2020) [hereinafter Huq, *Constitutional Rights*].

much damage will we really cause, and to whom? A machine might help us resolve this: a car armed with "artificial intelligence" and the right sensors might ascertain with something approaching near certainty the relevant facts, like the damage that would be caused to the one or the five given their physiques, and the car's size, speed, and road conditions. And if you think it relevant, such a machine might even determine the age and estimated quality-adjusted life-years remaining for each potential victim. Some machines—built on early models already in experimentation and relying on facial recognition—might even compare forecasts of each potential victim's charitable contributions, or whether they could have avoided the crash were *they* paying attention.[9]

Grim, but I'm not here to judge.

The problem is that, even with that knowledge and the benefit of time and computational power, I might still be uncertain about the right thing to do. Do I want the car to turn? Under what conditions? That is, I don't know what the right rule is, even if my self-driving car could perfectly execute it. I have what is called "normative uncertainty."[10]

So, what does normative uncertainty have to do with the "right to a human decision" debate? It turns out that normative uncertainty provides a reason we should flip the priority of arguments in this debate, in unexpected ways.

Currently in the debate, humanist "Arguments from Explanation" appear to be a strong point—and possibly the strongest point—of resistance against machinists' use of "Better Decision Arguments."

Arguments from Explanation, grounded in values like transparency, counsel against the use of machines (or certain types of machines) where the machine's decision cannot be explained.[11] These values matter because they are important to both auditing and justifying decisions.[12] In high stakes

---

9. Danielle Keats Citron & Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1, 4 (2025).

10. *See* Cox, *The Uncertain Judge*, *supra* note 4, at 751–54.

11. *See* Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1090–91 (2018); Daniel J. Solove & Hideyuki Matsumi, *AI, Algorithms, and Awful Humans*, 92 FORDHAM L. REV. 1923, 1935–38 (2024) (explaining how opacity can generate "automation bias"); Citron, *supra* note 1, at 1292–98 (discussing importance of judicial review); *see also* Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829, 1830 (2019) (arguing judges should demand explanations, thus "shaping the nature and form of xAI").

12. *See* Selbst & Barocas, *supra* note 11, at 1118–26 (discussing value of explanations); Deeks, *supra* note 11, at 1840–41; Kaminski & Urban, *supra* note 2, at 1979–80 ("Contestability relies on transparency[.]"); Pauline T. Kim, *Data-Driven Discrimination at*

contexts, like court decisions, it is important for the legal subject to understand not just the justification offered, but that the justification aligns with how the decision was reached—the actual operative "reasons" of the machine. And so, Arguments from Explanation demand not just a technical specification sheet or a justification, but an alignment between the two.[13] That is, what is wanted is an "aligned explanation."[14]

Humanists' Arguments from Explanation appear increasingly strong because they turn on technical limitations: the inscrutability and fundamental nonintuitiveness of modern machine learning technology.[15] While machines have long threatened transparency, it was usually by choice (e.g., secrecy).[16] But as machines grow more sophisticated, technical limits on the ability to explain machine decisions seem to render machines (and/or the humans who design them) *unable* to provide the kind of aligned explanations that are owed in many high-stakes decision-making contexts, like judging.[17] And so, the thought goes, if anything can ground a right to a human decision in such contexts, it is the right to an explanation.[18]

Work, 58 WM. & MARY L. REV. 857, 881, 921–23 (2017); Citron, *supra* note 1, at 1298; *see also* Nina Varsava, *Professional Irresponsibility and Judicial Opinions*, 59 HOUS. L. REV. 103, 118–20 (2021) (discussing related concerns with respect to judicial opinions).

13. *See* Coglianese & Lehr, *supra* note 3, at 39 n.151; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 639, 653–56 (2017) (noting inadequacy of technical transparency).

14. *See infra* Section I.A; Selbst & Barocas, *supra* note 11, at 1090–91 (describing nonintuitiveness as a barrier to explainability); Coglianese & Lehr, *supra* note 3, at 47–49.

15. *See* Selbst & Barocas, *supra* note 11, at 1091–96 (defining these terms).

16. *See, e.g.*, State v. Loomis, 2016 WI 68, ¶51, 371 Wis. 2d 235, 881 N.W.2d 749 (noting that developer of COMPAS recidivism calculator, citing trade secrecy, "does not disclose how the risk scores are determined"); Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1272 (2020) ("The primary obstacle to transparency is the pervasive practice of invoking trade secrecy[.]"); Selbst & Barocas, *supra* note 11, at 1091–93; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343, 792–95 (2018).

17. Selbst & Barocas, *supra* note 11, at 1091–96 (explaining that modern machine learning models are also opaque due to inscrutability and nonintuitiveness). *But see* Aniket Kesari et al., *Explaining Explainable AI* (Ctr. for Law & Econ., Working Paper No. 9, 2024), https://ssrn.com/abstract=4972085 (describing recent technological advances in "explainable AI" techniques).

18. *See, e.g.*, Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249, 1280–81 (2017); *cf.* Amin Ebrahimi Afrouzi, *John Robots*, *Thurgood Martian, and the Syntax Monster: A New Argument Against AI Judges*, 37 CAN. J.L. & JURIS. 369, 378–88 (2024). *But see* Andrew D. Selbst, Response, *A Mild Defense of Our New Machine Overlords*, 70 VAND. L. REV. EN BANC 87, 101–02 (2017) (arguing against "innate humanness of explanation").

I argue that Arguments from Explanation do not count in favor of humans in these contexts—and not because the human mind is a "black box" or human judges, like all humans, sometimes conceal illicit reasons, like bias or partisan preference.[19]

Rather, I argue that human explanations are incomplete and unaligned *even when* humans know their operative reasons and those reasons are good ones—especially in contexts where explanations are owed.[20] For example, judicial opinions often lack explanatory alignment and completeness where human judges aim at what they ought to do, are uncertain about what they ought to do, and respond to that uncertainty rationally.[21] Accordingly, whatever the merits of the right to an explanation, to the extent it seeks a kind of explanatory alignment, it does not provide a reason to favor even a well-behaved human judge over a machine.[22]

And so normative uncertainty seems to give the machinists a win.[23]

But what normative uncertainty gives with one hand, it takes with another.[24] This is because normative uncertainty calls into doubt the machinists' Better Decision Arguments—arguments that all humanists' principles really show is a right to a "better" decision, human or machine.[25] Alas, Better Decision Arguments make two mistakes.[26]

---

19. For arguments of this flavor and responses, see, for example, Brożek et al., *supra* note 2, at 430; Solove & Matsumi, *supra* note 11, at 1939; Boris Babic & I. Glenn Cohen, *The Algorithmic Explainability "Bait and Switch"*, 108 MINN. L. REV. 857, 885 (2023); Jenna Burrell & Marion Fourcade, *The Society of Algorithms*, 47 ANN. REV. SOCIO. 213, 222–23 (2021); Cary Coglianese & Lavi M. Ben-Dor, *AI in Adjudication and Administration*, 86 BROOK. L. REV. 791, 791 (2021); Cass R. Sunstein, *Governing by Algorithm? No Noise and (Potentially) Less Bias*, 71 DUKE L.J. 1175, 1177–78 (2022) (arguing that, unlike humans, algorithms can eradicate "unwanted variability in judgments").

20. *See infra* Section I.B.

21. Also of note: it doesn't turn on contested views about judges being permitted to conceal their reasons. *See* Courtney M. Cox, *Super-Dicta*, 173 U. PA. L. REV. 1575, 1600 (2025) [hereinafter Cox, *Super-Dicta*].

22. *See infra* Section I.B.

23. *See infra* Section I.

24. *See infra* Section II.

25. Or a "better together" machine/human hybrid decision. *See generally* Huq, *Human Decision*, *supra* note 2 (rejecting a general right to purely human decision-making and instead proposing a "*right to a well-calibrated machine decision* that folds in due process, privacy, and equality values" while emphasizing that the relevant question is timing and structure of human involvement); Coglianese, *supra* note 2 (labeling Huq's argument as about the "right to better decisions"); Frank Fagan & Saul Levmore, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. L. REV. 1, 3–7, 13–28 (2019) (providing framework for "the Human-AI combination"); Solove & Matsumi, *supra* note 11, at 1923 (arguing that "Better Together Arguments" underestimate the difficulty of combining forces);

First, Better Decision Arguments assume that we—the engineers, policymakers, and end users—know what better is.

Second, Better Decision Arguments assume that when we don't know what better is, the solution is to just pick. And instead of picking in the moment, what's wrong with picking an algorithm or objective function in advance? It might even be better to leverage the algorithm as a "pre-commitment device," especially in contexts like judging.[27]

The first mistake is obvious: the problem of normative uncertainty has been missed. This is not surprising. This problem of normative uncertainty is often confused with other difficulties.[28] There is the technical problem of value alignment, of getting the machine to do what we want it to do.[29] And there is the political problem of disagreement, of how to resolve differing views about what the machine should do.[30] But normative uncertainty lurks behind both of these problems: Even if we could resolve problems of descriptive uncertainty, disagreement, and value alignment, we might still be unsure what we want the car to do—and not just in situations we haven't imagined, but also in ones we have.[31]

Missing the problem of normative uncertainty often leads to the second mistake: to assume that we should just pick—or identify the right political process for resolving disagreement about what to pick. But as I will argue, that obvious solution—the standard AI paradigm—is a nonstarter.[32] It both violates dominance and fails to take the stakes into account.[33]

So, to recap, normative uncertainty seems to undermine leading strategies on both sides. Arguments from Explanation do not count in favor

---

see also Rebecca Crootof et al., *Humans in the Loop*, 76 VAND. L. REV. 429, 467–73 (2023) (arguing that simplistic "Better Together" approaches are a "trap").

26. *Cf.* Cox, *The Uncertain Judge*, *supra* note 4, at 765–86 (describing why normative uncertainty is overlooked in judging).

27. Huq, *Human Decision*, *supra* note 2, at 675 (noting that "encoded judgments" work like opinions, serving "as a pre-commitment to generality and as a safeguard against personalistic or arbitrary state action").

28. *See* Cox, *The Uncertain Judge*, *supra* note 4, at 765–86 (explaining difficulty of isolating problem in judicial context).

29. *See infra* Section II.B.1.

30. *See infra* Section II.B.1.

31. *See infra* Section II.B.1.

32. *See infra* Section II.B.2; *see also* Cox, *The Uncertain Judge*, *supra* note 4, at 776–86; STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 4–5 (4th ed. 2021) ("The standard model has been a useful guide for AI research since its inception, but it is probably not the right model in the long run.").

33. *See infra* Section II.B.2; *see also* Cox, *The Uncertain Judge*, *supra* note 4, at 776–86; MACASKILL ET AL., *supra* note 4, at 40–41; LOCKHART, *supra* note 4, at 4.

of human decisionmakers, like judges, in contexts where explanations for decisions are owed.[34] And Better Decision Arguments commit critical errors in both ignoring the problem and offering a nonstarter of a solution.[35]

Where does this leave us? We are now positioned to discuss more seriously the implications of undermining the humanist's Arguments from Explanation and the machinist's Better Decision Arguments.[36] The core question is not really about humans v. machines, but about the timing and manner of human involvement in decisions in particular cases.[37] But what should inform that choice?

Some scholars have suggested that when individual-level decisions have "shades" of "normativity," humans are needed.[38] Those arguments had been predicated, in part, on technical limitations—machines only recently became capable of imitating moral reasoning.[39] These "normativity"

---

34. *See infra* Section I.B.

35. *See infra* Section II.B.

36. Obviously, I have been taking some liberties in describing a "two-sided" all-or-nothing debate for sake of exposing the problem with the argumentative strategies on both sides. Everyone knows that "[i]t's humans all the way down," and that the real question is about the timing and manner of human involvement in different contexts. Crootof et al., *supra* note 25, at 443 (collecting literature) ("We are not the first to note this."); Huq, *Human Decision*, *supra* note 2, at 650 ("[I]f it is not meaningful to speak of machine decisions that do not have a human in the loop, there is a question as to why the timing of necessary human involvement makes a practical difference."); Kaminski & Urban, *supra* note 2, at 1972 ("[The issue is] [w]ho makes decisions and when decisions are made."). I will return to this in Section III.A. Although I use Professor Huq's argument, and its reception by others, as the prime example of the Better Decision Argument, he believes institutional reform will also be important to deciding when and whether machine—at least as I read him. *See* Huq, *Human Decision*, *supra* note 2.

37. This question is the important question for lawmakers and regulators. *See* Huq, *Human Decision*, *supra* note 2, at 615, 620–28 (explaining why this question matters to lawmakers and canvassing law that might ground a right to a human decision).

38. *Id.* at 686.

39. Harry Surden, *Artificial Intelligence and Law—An Overview of Recent Changes: Keynote Address at the 2024 IRA C. Rothgerber Jr. & Silicon Flatirons Conference on Artificial Intelligence and Constitutional Law*, 96 U. COLO. L. REV. 375, 380 (2025) ("[P]rior to 2022 such AI capabilities were not remotely possible at the current level of usefulness and sophistication that we see today."); *cf.* James Grimmelmann et al., *Generative Misinterpretation*, 63 HARV. J. LEGIS. 229, 232–33 (2026) ("These LLM proponents['] . . . bottom lines are broadly similar: LLMs are already 'good enough,' and judges should seriously consider trusting them to assist with interpretative work in actual cases. . . . We respectfully dissent.").

arguments thus now seem poised to succumb to Better Decision Arguments.[40]

But Better Decision Arguments get it wrong, and in a way that informs the relevant question to be asked. The critical question is not whether machines are "better" or more "accurate" than humans in making these decisions. The critical question is: what is the appropriate scope of choice for resolving our normative uncertainty in this space? Is the appropriate scope of choice an algorithm, a rule, stretched across a number of cases? Or is it some smaller unit, case-by-case, or issue-by-issue?

The problem of scope of choice is terribly, terribly difficult. Different answers can lead, paradoxically, to conflicting results.[41]

And so here, we can see one of the reasons for the seeming intractability and discomfort within the human decision debate. It is not *merely* that machines crystalize questions of normative uncertainty and that the standard frame runs into rational errors.[42] It is that the debate intersects with the problem of normative uncertainty at *the problem's most difficult joint.*

A pit in one's stomach about machines? No wonder.

I do not know the answer to this scope-of-choice question writ large. But there are reasons to favor different scopes of choice in different contexts. In self-driving cars, the point is to avoid trolley problems—to avoid the situation in the first place—and when it can't be avoided, there's not time in the moment to debate; even human drivers use rules of thumb.[43] And so, a broader scope of choice for self-driving cars is probably both appropriate and unavoidable.[44] By contrast, the point of adjudication may well be to run straight into the hard cases, not avoid them, and to come out the other side.[45]

---

40. *See* Kiel Brennan-Marquez & Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. CRIM. L. & CRIM. 137, 142 (2018) (searching for alternative grounds for why human decisions matter in light of this possibility).

41. Lewis A. Kornhauser & Lawrence G. Sager, *The One and Many: Adjudication in Collegial Courts*, 81 CALIF. L. REV. 1, 11–17 (1993) (discussing doctrinal paradoxes on collegial courts); *see also* Christian List & Philip Pettit, Response, *On the Many as One: A Reply to Kornhauser and Sager*, 33 PHIL. & PUB. AFFS. 377, 382–83 (2005) ("The integrity challenge . . . arises for any set of nontrivially interrelated propositions.").

42. *See infra* Part II.

43. *See infra* Section III.A.

44. *See infra* Part III.

45. *Cf.* Seana Valentine Shiffrin, *Inducing Moral Deliberation: On the Occasional Virtues of Fog*, 123 HARV. L. REV. 1214, 1217 (2010) ("[T]his sort of induced moral deliberation is important for our moral health and for an active, engaged democratic citizenry."); Rebecca Stone, *Rights, Remedies, and Normative Uncertainty About Justice*, 31(1) LEGAL THEORY 114, 114 (2025) (arguing that remedial law should be used to facilitate deliberation between the parties about what justice requires); Aditi Bagchi, *Private Law and Public Discourse*, 65 ARIZ.

By focusing on normative uncertainty, we can see more clearly why the *mode* of decision-making—human or machine—matters, and have another tool in our arsenal for making an informed choice about when to deploy machines.[46] We will also discover a new and unexpected use case for machines to assist with normative judgments.[47] And so, even if I have replaced one difficult question with another, we have at least made progress.

The contributions of this paper are both timely and timeless. Could this paper have been written before? Absolutely. It joins the tradition of using machines to think through and engage with difficult questions of jurisprudence, and vice-versa.[48] But it also seeks to enter that timeless debate at a moment when technology—both its strengths and its limitations—seem to have distorted the relative prioritization of considerations for and against humans.

Perhaps more importantly, the normative debate needs to catch up with the technical problems. Engineers have been searching for answers about how to ensure value alignment in the shadow of normative uncertainty, but without the language to understand the problem.[49] Without that understanding, and with normative debates seemingly at an impasse, some computer scientists have turned to crowd-sourcing "public morality" to

---

L. REV. 541, 576 (2023) (arguing that private law disputes form a critical part of reaching consensus on moral and political norms).

46. *See infra* Section III.A.

47. *See infra* Section III.B.

48. *See, e.g.*, Trevor Bench-Capon et al., *A History of AI and Law in 50 Papers: 25 Years of the International Conference on AI and Law*, 20 A.I. & L. 215 (2012); DOUGLAS WALTON, ARGUMENTATION METHODS FOR ARTIFICIAL INTELLIGENCE IN LAW (1st ed. 2005); NEIL MACCORMICK, INFORMATICS AND THE FOUNDATIONS OF LEGAL REASONING 99–117 (Zenon Bankowski et al. eds., 1st ed. 1995); *see also, e.g.*, Lawrence B. Solum, *Artificial Meaning*, 89 WASH. L. REV. 69, 83–85 (2014); *cf.* Michael J. Madison, *Fair Play: Notes on the Algorithmic Soccer Referee*, 23 VAND. J. ENT. & TECH. L. 341 (2021) (examining how use of video review "affects the law and related governance of soccer" as a case study for thinking about how computational tools affect judging and judgment). And of legal questions to think through difficult questions about machines, see generally Lawrence B. Solum, *Legal Personhood for Artificial Intelligences*, 70 N.C. L. REV. 1231 (1992).

49. *See, e.g.*, Edmond Awad et al., *The Moral Machine Experiment*, 563 NATURE 59, 59–64 (2018) ("Car manufacturers and policymakers are currently struggling with these moral dilemmas . . . .").

justify their design choices.[50] They also have been working on "uncertain AI" as a way of addressing the value-alignment problem.[51]

In short, the standard framing of the human decision debate as centered around Better Decision Arguments has missed that the technology has moved on.[52] Engineers have figured out that they should not "just pick," that perhaps the machine should say "I don't know."[53] But the engineers are not entirely clear on the extent of the problem or how to solve it in a principled way. To get there, humans need to start saying "I don't know" back to the machines. We need to correctly identify the problem before we can work together on principled solutions.

This Article proceeds as follows: Part I explains why humanist Arguments from Explanation do not counsel in favor of humans, at least in high stakes contexts where explanations for decisions are owed. Part II explains why machinist Better Decision Arguments commit critical errors in arguing that the relevant limitations are technical. Part III brings these arguments together to refine the questions underlying the debate and to offer a path forward.

I. THE ILLUSION OF EXPLANATORY ALIGNMENT

My argument about how normative uncertainty should reorient the human decision debate begins with arguments from transparency and explanation—what I will call "Arguments from Explanation." These types of arguments are commonly identified as grounding a right to human decisions in high-stakes contexts where explanations are owed. [54]

---

50. *See, e.g.*, *id.* ("As a response to these challenges, we designed the Moral Machine, a multilingual online 'serious game' for collecting large-scale data on how citizens would want autonomous vehicles to solve moral dilemmas in the context of unavoidable accidents.").

51. RUSSELL & NORVIG, *supra* note 32, at 34 (describing new model).

52. *Id.*

53. *Id.* at 33–34 ("[W]e want machines that strive to achieve human objectives but know that they don't know for certain exactly what those objectives are."); *see* Andreia Martinho et al., *Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence*, 31 MINDS & MACHS. 215, 233 ("[W]e believe that, in general terms, the decisions made by a morally uncertain AI (equipped with a latent class choice model of morality) should be preferred to the decisions made by an AI that is morally certain.").

54. Huq, *Human Decision*, *supra* note 2, at 656; *see also* Brennan-Marquez, *supra* note 18, at 1300 ("Explanations are what allow us to answer [whether some are subjected to police searches] consistently with our values."); Ray Worthy Campbell, *Artificial Intelligence in the Courtroom: The Delivery of Justice in the Age of Machine Learning*, 18 COLO. TECH. L.J. 323, 342–43 (describing the concerns over use of COMPAS—whose "proprietary algorithm" cannot be reviewed for "accuracy and fairness"—to predict recidivism risk); Selbst & Barocas, *supra*

My goal is not to theorize existing *legal* rights to explanations, though a common machinist strategy in these debates is to argue that "explainable AI" techniques can satisfy those legal rights.[55] Rather, my target is the normative or pre-legal principle that provides a justification for such formalized rights and, more broadly, for a right to a human decision.[56]

To get my argument off the ground, I will map, roughly, what is at stake in the demands for an explanation and calls for "transparency." I also will map why and how recent advances in AI seem to strengthen the claim that machines—at least, those which are most likely to be used—cannot offer the kind of explanation which Arguments from Explanation demand. This is the work of Section A.

Then, in Section B, I explain how normative uncertainty undermines the availability of such explanations by human decisionmakers, even—and especially—when human decisionmakers act as they should. As a result, whatever their merits, Arguments from Explanation do not count in favor of human decisionmakers—at least not on the ground that human decisionmakers can offer what machines cannot.

A. The Right to an (Aligned) Explanation

At present, some of the strongest arguments in favor of human decisionmakers appear to be arguments from "transparency" and the "right to an explanation"—Arguments from Explanation.[57] But what are "transparency" and "explanation"? That is, what do these arguments

---

note 11, at 1118 ("There are several reasons to view explanation as a good unto itself, and perhaps a necessary part of a system constrained by law, including a respect for autonomy, dignity, and personhood.").

55. *E.g.*, Coglianese & Lehr, *supra* note 3, at 6. Some scholars define "explainability" in terms of such techniques. *See, e.g.*, Kate Vredenbergh, *Transparency and Explainability for Public Policy*, LSE PUB. POL'Y REV., Nov. 4, 2024, art. 4, at 1, 3 (2024) ("[A]n explainable model is one where an expert can only understand how a model produced its output if they use additional tools, such as techniques from machine learning to generate explanations of the model." (citing Emily Sullivan, *Understanding from Machine Learning Models*, 73 BRIT. J. FOR PHIL. SCI. 109, 109 (2022))).

56. My strategy thus follows that of Professor Huq. *See* Huq, *Human Decision*, *supra* note 2, at 615 ("My focus is not on the form of such a right . . . or how it is implemented . . . but more simply on what might ab initio justify its creation.").

57. Or at least, most popular! *See* Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 262 (2019) ("Perhaps the most widely appreciated risk of Al decision-making is that it could function in ways that are hard or impossible for humans to comprehend."); Lehr & Ohm, *supra* note 2, at 705 (describing "an onslaught of work on . . . 'explainability'").

demand from machines? Why do they demand it? And why are machines—especially those powered by deep learning technology—thought unable to provide it?

My interest in these questions is to identify what it is that Arguments from Explanation, charitably read, are after—what is it that humanists want that machines (putatively) can't provide?—so that we might consider whether human explanations can offer it. Accordingly, I do not evaluate the merits of these arguments on their own terms, or whether their assessment of technical limitations is correct. Nor do I address how they relate to the appropriate design and interpretation of related legal regimes.[58] Similarly, I will not attempt to reconcile competing definitions of "transparency," "explanation," and "explainable AI."[59]

My goal is just to identify the core of what is demanded—what is valued—that is thought to count against machines and in favor of humans.

While I'm not interested in a taxonomy of the different types of explanation, I'll note at the outset that different contexts may require different kinds and levels of explanation.[60] Specifically, some of these contexts are less demanding than others.[61] And so, the demands within some of these contexts may be more readily met by machines.

My strategy will thus be to focus on the humanist's argumentative strategy where it might appear strongest: the high stakes and explanatorily

---

58. There is much disagreement about how to interpret these legal regimes—both about what the regimes require by way of disclosure and whether the required disclosures would satisfy the normative ideals of transparency and explainability. *See, e.g.*, Finale Doshi-Velez et al., Accountability of AI Under the Law: The Role of Explanation 1, 3 (Dec. 22, 2019) (unpublished paper) (on file with the Berkman Klein Ctr. for Internet & Soc'y), https://www.ssrn.com/abstract=3064761 [https://perma.cc/D7B5-3BHR]; Adrien Bibal et al., *Legal Requirements on Explainability in Machine Learning*, 29 A.I. & L. 149, 164–67 (2020); Margot E. Kaminski, *The Right to Explanation, Explained*, 34 BERKELEY TECH. L.J. 189, 192–93 (2019) ("Arguments over the purported right to explanation obscure the true substance and depth of the GDPR's algorithmic accountability regime."); Sandra Wachter et al., *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J. L. & TECH. 841, 844–53 (2017); Selbst & Barocas, *supra* note 11, at 1099. *See generally* Citron, *supra* note 1 (offering framework for technological due process).

59. *E.g.* Coglianese & Lehr, *supra* note 3, at 32–33 (distinguishing "fishbowl transparency" from "reasoned transparency"); Selbst & Barocas, *supra* note 11, at 1090–91 (discussing distinction between "transparency" and "explainability"); Kesari et al., *supra* note 17, at 22 (identifying different types of explanation offered by xAI technologies); Kroll et al., *supra* note 13, at 657, 662–72 (describing alternatives to "transparency" for "verifying procedural regularity").

60. *See* Kate Vredenburgh, *The Right to Explanation*, 30 J. POL. PHIL. 209, 225 (2022); Kesari et al., *supra* note 17, at 22; Deeks, *supra* note 11, at 1840–41.

61. *See* Kesari et al., *supra* note 17, at 22; Selbst, *supra* note 18, at 97–98.

demanding context of adjudication. It is the quintessential example of a context where explanations are owed, explanations of a kind that machines are thought peculiarly unfit to offer.[62]

So, what do humanists demand when they demand "transparency" or an "explanation"?

One way to think about transparency and explanation is in terms of the questions asked.[63] These questions will vary depending on who asks them.[64] For example, the public may want to know what the technology is, why it was adopted, and that steps were taken to ensure its performance; regulators may want to know the answer to similar questions, but at a greater level of detail.[65] And the subjects of the decisions will often want to know the basis for particular decisions—for purposes of challenging the decision in their own case, for comparison of their case to that of others, and for guidance going forward.[66]

Although these questions are asked at different levels of generality, and so demand different information in response (e.g., more or less technical detail), they generally go to the same two basic questions. Those basic questions are: "Why is that the decision?" and "How do I know that it is right?" Together, I suggest, the questions call for an "aligned explanation."

To see this, note that the first basic question—"Why is that the decision?"—calls for two types of explanation.

In one register, the question calls for a description: what features and what logic were used to reach the decision? Call the answer to such a question "explanation-as-description," and the features and logic the "operative reasons."[67]

---

62. *E.g.*, Afrouzi, *supra* note 18, at 374–76; Babic & Cohen, *supra* note 19, at 859–60; Deeks, *supra* note 11, at 1840–41; Huq, *Human Decision*, *supra* note 2, at 640; Huq, *Constitutional Rights*, *supra* note 8, at 1897.

63. *See* Selbst, *supra* note 18, at 97 ("[A] desire for an 'explanation' in general is underspecified and only makes sense when offered in opposition to 'no explanation.'").

64. *See* Kesari et al., *supra* note 17, at 7–8.

65. *Id.*; W. Nicholson Price II, *Medical AI and Contextual Bias*, 33 HARV. J. L. & TECH. 65, 99 (2019) (highlighting challenges of translating medical AI from one context to another and arguing that AI "opacity makes it hard to spot the problems of contextual bias").

66. *See* Kesari et al., *supra* note 17, at 14–17; *see, e.g.*, Citron, *supra* note 1, at 1252–98.

67. I borrow the term "operative" from Scanlon, though not all of his machinery (pun intended), T. M. SCANLON, WHAT WE OWE TO EACH OTHER 19 (Angela Smith ed., 1st ed. 1998), because it seems more appropriate to the machine case than the more commonly used phrase "motivating reasons." The phrase "motivating reasons" sometimes invokes psychological states, like belief and intention and desire, that are not relevant to the machine case. *Cf.* Pamela Hieronymi, *Reasons for Action*, 111 PROC. ARISTOTELIAN SOC'Y 407, 411 (2011) (collecting literature) (describing the conventional view of "motivating" or "operative"

In another register, the question calls for a justification: why should I accept the decision—based on what features and by what logic, is the decision justified? Call the answer to such a question "explanation-as-justification," and the features and logic offered by such an explanation that (putatively) supports the connection between inputs and outputs the "justificatory reasons."[68]

I obviously indulge in a bit of anthropomorphizing by using "reasons," and especially "operative reasons," but a technical description is clumsier and risks becoming obsolete, and my main interest anyways will be to argue that you don't get what you want with respect to reasons in the human case, either.

The second question—"How do I know that it is right?"—calls for information that would verify the answers to the first set of questions. The asker wants enough technical detail to confirm that the explanation-as-description is accurate—that the machine functions as described and did not malfunction in the individual's case. And the asker wants enough descriptive detail about the features and logic that were actually used—that is, the asker wants an explanation-as-description—to know whether the explanation-as-justification holds water and is not hiding reliance on illicit reasons: that the reasons proffered as justifying the decision *are in fact* the reasons why the decision was made.

In each of these cases, this set of questions implies that a kind of "explanatory alignment" is sought (or presumed).[69] If the explanation-as-

---

reasons as "considerations that someone took to count in favour of an action, whether or not they actually count in favour of it—those considerations someone *treated as* 'normative' reasons" before adopting a modified view related to psychological facts); Vredenburgh, *supra* note 60, at 211 (using "motivating reasons" to refer to "reasons on which [a person] act[s] that they take to count in favor of their action"); Bernard Williams, *Internal and External Reasons*, *in* RATIONALITY IN ACTION: CONTEMPORARY APPROACHES 387, 387–88 (Paul K. Moser ed., 1990) (arguing that an individual's reasons for actions are "relative" to their internal, subjective motivations); Babic & Cohen, *supra* note 19, at 867–70 (explaining the difference between "interpretable AI/ML," where "an ordinary person" can understand how the model's features relate to the model's prediction, and "explainable AI/ML," where the behavior of an uninterpretable AI/ML is approximated by an interpretable model); Mathilde Cohen, *Sincerity and Reason-Giving: When May Legal Decision Makers Lie*, 59 DEPAUL L. REV. 1091, 1097 n.23 (2010) ("The criterion for a motivating reason is roughly that it explains rather than justifies a person's actions or decisions.").

68. I do not use the phrase "normative" to distinguish these two types of reasons, because the "operative" reasons are, in a sense, normative, at least in the human case: they are "reasons." *Cf.* Vredenburgh, *supra* note 55, at 6–8.

69. For an example of an argument identifying alignment as the object of arguments from explanation, see generally Babic & Cohen, *supra* note 19 (using term "sincerity").

justification does not align with the explanation-as-description, it will fail to provide guidance: future behavior that is similar with respect to justificatory features may be subject to a different decision if it differs with respect to the operative features relied upon. And relatedly, if the explanation-as-justification fails to align with the explanation-as-description, then there may be a basis for challenging the decision.[70]

For example, the explanation-as-justification may be sound as far as it goes, but if there is no explanatory alignment, then I lack reassurance that I have been treated fairly: Someone whose justificatory features are the same as mine, but whose operative features differ, may receive a different, more favorable decision. Like cases are not treated alike.

Similarly, if the explanation-as-description does not align with the technical underpinnings, then it would seem I lack reassurance that the explanation-as-description—what I have been given as the "operative" reasons—are in fact the operative reasons. There has been a technical error in the functioning of the machine, at least as described.

The demand for "explanatory alignment"—or an aligned explanation—is most easily illustrated by considering a case where a decisionmaker relies on *inappropriate* or *illicit* reasons.[71] For example, if a judge were to always rule against a protected group, or in favor of attractive women, or based on his opinion of the person's attire, then even if the explanation-as-justification of the decision held water, there would be grounds to challenge the decision because it failed to align with the judge's operative reasons.[72]

Arguments from Explanation seem best positioned to ground a right to a human decision because it is not clear that machines can provide—or

---

70. *See* James Grimmelmann & Daniel Westreich, *Incomprehensible Discrimination*, 7 CALIF. L. REV. ONLINE 164, 173–74 (2017) (narrating fictional opinion applying nondiscrimination law to require companies using algorithms in hiring "to build a public record establishing that its algorithms work as described and for the right reasons").

71. Kim, *supra* note 12, at 881, 921–23 ("This lack of transparency makes it difficult to know if any observed bias is simply a byproduct of justifiable business considerations or the result of flaws in the model's construction.").

72. The ability to succeed in such showings is another thing. *See, e.g.*, Robert P. Schuwerk et al., *Disqualification—Bias*, *in* 48C HANDBOOK OF TEXAS LAWYER AND JUDICIAL ETHICS § 40:13 (2025) ("[O]ver 20 years ago Texas case law began to posit a ground for *disqualification* with a certain degree of bias or prejudice."); *see also* Emily Berman, *A Government of Laws and Not of Machines*, 98 B.U. L. REV. 1277, 1319 (2018) ("If a juror expresses racial prejudice . . . a verdict might be overturned." (citing Pena-Rodriguez v. Colorado, 580 U.S. 206, 206 (2017))).

developers disclose—an explanation that meets the demand for explanatory alignment.[73]

Why can't machines provide the requisite explanation? The problem is the technology, or so the argument goes. While developers and their machines have long generated obfuscation by choice—frequently under the guise of trade secrecy—more advanced technologies may be opaque even to their creators who have the requisite technical expertise.[74] This new opacity has several sources, depending on the technology in question.

The first source is inscrutability: "a situation in which the rules that govern decision-making are so complex, numerous, and interdependent that they defy practical inspection and resist comprehension."[75] Inscrutability renders explanation-as-description unavailable: it is not possible, even for those with access and expertise, to understand the machine's operative reasons. And if there is no way to understand the machine's operative reasons—if an explanation-as-description is unavailable—then there is no way to verify that any proffered justificatory reasons align with the machine's operative reasons.

Inscrutability has long been recognized as a strong basis for arguments from explanation.[76] Even before the advent of modern machine-learning technologies, algorithms could be sufficiently complex as to render the machine's operative reasons inscrutable, and an aligned explanation thereby unavailable.[77]

But as predictive technology improves, arguments from explanation have an even firmer basis: a machine's operative reasons are nonintuitive. That is, even where the machine's operative reasons can be understood—where "the statistical relationship that serves as the basis for decision-making

---

73. *See* Afrouzi, *supra* note 18, at 374–76; Babic & Cohen, *supra* note 19, at 883. Babic and Cohen do not argue in favor of an absolute right to a human decision, but instead argue that the lack of alignment should inform which decisions should be reserved for humans rather than machines, and when machines, if used, should be limited to types that can provide the necessary alignment (even if at the cost of predictive accuracy). Their treatment of the human comparator differs from mine: I give humans more credit, and they still come up wanting. But that is appropriate given their primary focus, which is selection of the appropriate type of technology in light of the interpretability of that technology. Babic & Cohen*, supra* note 19, at 908–09.

74. *See* Selbst & Barocas, *supra* note 11, at 1093–94; *see also* Wexler, *supra* note 16, at 1421 (evaluating use of trade secrecy to protect proprietary algorithms).

75. Selbst & Barocas, *supra* note 11, at 1094 (defining "inscrutability").

76. *Id.*

77. *See id.*

might be readily identifiable"[78]—the operative reasons "may defy intuitive expectations about the relevance of certain criteria to the decision."[79]

This lack of intuitiveness severs the alignment between the explanation-as-description and the explanation-as-justification, largely by rendering an explanation-as-justification unavailable. We can understand the explanation-as-description—what the operative reasons of the machine *are*. But we fail to grasp how those operative reasons could align with any possible picture of the justificatory reasons for believing a prediction or making a decision, beyond taking on faith that the operative reasons somehow track the desired output.[80]

Although Selbst and Barocas argue that "[t]he demand for intuitive relationships is not the demand for disclosure or accessible explanations,"[81] but a dispute about the appropriate bases for a decision, the nonintuitiveness of the explanation-as-description also supports an argument *from explanation* to the extent that the explanation demanded must offer something more than "take it on faith" that an explanation-as-justification is available if only we could understand it. Reasons and explanations are two sides of the same coin in the context of justification.[82]

Indeed, depending on the tightness of alignment demanded, there may be a further problem, namely, that any attempt to overcome the above problems through "explainable AI" or "xAI" technologies—technologies that attempt to explain a base model in more accessible terms by, essentially, creating a model of the model—will necessarily fail to exhibit alignment.[83] Or so the argument goes.[84] This is because the "explanation" provided by xAI is necessarily *not* identical to the model.[85] Either the model is interpretable—i.e., does not exhibit inscrutability or nonintuitiveness—or else it is not. And if it is not, the best available "explanation" is always a

---

78. *Id.* at 1097.

79. *Id.*; *see also id.* at 1091 n.30 ("We intentionally use the term 'nonintuitive' rather [than] the word 'unintuitive' or 'counterintuitive.' In our view, 'unintuitive' implies a result that would not be expected but is easily understood once explained, and 'counterintuitive' suggests a phenomenon that is opposite one's expectations. Instead, we intend to refer to a phenomenon about which intuitive reasoning is not possible.").

80. *Id.* at 1097.

81. *Id.*

82. *Cf.* Hieronymi, *supra* note 67, at 421 ("Reasons, then, are considerations that bear or are taken to bear on questions."); Frederick Schauer, *Giving Reasons*, 47 STAN. L. REV. 633, 636 (1995) (noting relationship between explanation and justification).

83. Babic & Cohen, *supra* note 19, at 864.

84. *Id.*

85. *Id.* at 869–70 (providing technical explanation).

post-hoc heuristic that merely *approximates* the model. After all, if the xAI "explanation" *matched* the underlying model, then the xAI explanation would also be inscrutable and/or nonintuitive.[86]

To recap where we are to this point: Arguments from Explanation in favor of human decisions, and against machine ones, generally demand an aligned explanation for decisions: one in which the answers to the questions "Why this decision?" in both the descriptive and justificatory sense align.

Machines cannot provide such an explanation because it is precisely this alignment that "black box" technologies threaten. An explanation-as-description will not "make sense" as supporting an explanation-as-justification if either the explanation-as-description cannot be understood (the problem of inscrutability) or if the link between the operative reasons and the justificatory reasons is nonintuitive such that we need to take the availability of an explanation-as-justification on faith (the problem of faith or noninuitiveness).

Developers are thus left with a choice. Either switch to a technology that does not implicate either the problem of inscrutability or of faith, in which case you get the demanded explanation but not the sophistication necessary for the setting, like adjudication.[87] Or else, gain a technology that appears capable of performing the function (however it manages to do it), but without the explanation.

And so, it is thought, Arguments from Explanation count strongly in favor of a human decision where explanations of those decisions are important, but where the decision is sufficiently complicated that only "black box" technologies seem up to the task.

I want to emphasize that the foregoing discussion has sought only to identify one of the core concerns of Arguments from Explanation that are thought unavailable from machine decisions. I do not here seek to evaluate the strength of the argument in the context of machines.[88] It may be that,

---

86. For further discussion, see Babic & Cohen, *supra* note 19, at 893–94; Selbst & Barocas, *supra* note 11, at 1129.

87. *Compare* Huq, *Human Decision*, *supra* note 2, at 636 n.123 (deriding the prospect of automated judging as "fanciful"), *with* John Zhuang Liu & Xuevao Li, *How Do Judges Use Large Language Models? Evidence from Shenzhen*, 16 J. LEGAL ANALYSIS 235, 238–40 (2025) (discussing workflow where AI-generated judicial opinions are then revised by judges); *see also* Ross v. United States, 331 A.3d 220, 229–30 (D.C. 2025) (Howard, J., concurring) (same); Snell v. United Specialty Ins., 102 F.4th 1208, 1221–25 (11th Cir. 2024) (Newsom, J., concurring) (same).

88. For critical analysis, see Lehr & Ohm, *supra* note 2, at 705–10; Kroll et al., *supra* note 13, at 657, 662–72.

contra the critics, these concerns can be alleviated in the context of machines, that the technical limitations are not so dire after all, and that with time, we may come to appreciate the justificatory reasons tracked by the machine's operative reasons. I don't claim otherwise.

This is because my point is not about the machines. It is about the argument. Arguments from Explanation argue against machine decisions on the grounds that machines cannot provide aligned explanations. But for this argument to count in favor of humans, it must be the case that humans can provide aligned explanations. And it turns out that, at least in one critical context, they can't.

B. The Human Judge

To show how Arguments from Explanation do not count in favor of humans, I will consider an example in which explanations are thought critical: adjudication. The use of machines in this context—in a way that directly informs the decision—often inspires dread. For example, machine critics often invoke the example of the COMPAS recidivism calculator, which "predicts" the likelihood that a defendant will re-offend if released.[89] While there are many criticisms of this tool,[90] legal challenges frequently emphasize how the "black box" nature of the tool precludes explanatory alignment: defendants are unable to challenge whether the algorithm's features and logic—its operative reasons—connect with justificatory reasons for believing or relying on its output as accurate.[91]

---

89. Ben Green, *The Flaws of Policies Requiring Human Oversight of Government Algorithms*, COMPUT. L. & SEC. REV., 2022, at 1, 9–10, 16 (using Loomis and COMPAS as an example to argue that "[t]he assumption of effective human oversight provides a false sense of security in adopting algorithms"); Katherine Freeman, *Algorithmic Injustice: How the Wisconsin Supreme Court Failed to Protect Due Process Rights in* State v. Loomis, 18 N.C. J.L. & TECH. 75, 89–96 (2016); *see also* State v. Loomis, 2016 WI 68, ¶¶ 13–14, 371 Wis.2d 235, 881 N.W.2d 749 (explaining how COMPAS offers predictive scores along several metrics).

90. *See, e.g.*, Frederik J. Zuiderveen Borgesius, *Strengthening Legal Protection Against Discrimination by Algorithms and Artificial Intelligence*, 24 INT'L J. HUM. RTS. 1572, 1574–76 (2020); Campbell, *supra* note 54, at 342–43 (noting concerns about bias); *see also* Julia Angwin et al., *Machine Bias*, PROPUBLICA (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing [https://perma.cc/5TEY-ETJT].

91. *See Loomis*, 881 N.W.2d at 756–57 (ruling on challenge to COMPAS risk assessment as violating "a defendant's right to be sentenced based upon accurate information, in part because the proprietary nature of COMPAS prevents him from assessing its accuracy"); *see also* People v. Wakefield, 107 N.Y.S.3d 487, 494–95 (N.Y. App. Div. 2019) (rejecting defendant's argument that his inability to access the source code of algorithmic DNA matching software used to match his DNA to DNA at the crime scene violated his right to confront witnesses).

But it turns out that explanatory alignment is often not available in the human case, either. This isn't just the case when human judges behave badly, nor does it depend on arguments that minds are a "black box."[92] Nor does it turn on a view about the permissibility of concealing operative reasons—about whether judges may, as a matter of good jurisprudence, decline to offer a full explanation.[93]

Rather, explanatory alignment fails where humans are called on to explain their decisions but face normative uncertainty about what they ought to do and respond to that uncertainty rationally.[94] In particular, it turns out that when a judge responds rationally to normative uncertainty, the proffered explanation will lack explanatory alignment.[95]

At the outset, I want to emphasize that I am *not* arguing that it is a bad thing that you don't get alignment in the human case. Whether judicial opinions should exhibit alignment (or are even expected to) is a point about which there is some disagreement.[96] And there are interesting questions about whether, when a judge employs their "situation sense" and writes an

---

92. *See* Solove & Matsumi, *supra* note 11, at 1926 (criticizing machinist "Awful Human Arguments" as underestimating humans). For versions of the "Blackbox Brain" arguments and their criticisms, see, for example, Brożek et al., *supra* note 2, at 430.

93. *See, e.g.*, Cohen, *supra* note 67, at 1101–03 (arguing against sincerity qua alignment); Paul Butler, *When Judges Lie (and When They Should)*, 91 MINN. L. REV. 1785, 1785–86 (2007) (citing Robert M. Cover, Book Review, 68 COLUM. L. REV. 1003, 1005–08 (1968)); ROBERT M. COVER, JUSTICE ACCUSED: ANTISLAVERY AND THE JUDICIAL PROCESS 119–22 (1975).

94. Cox, *Super-Dicta*, *supra* note 21, at 1600–17; Courtney M. Cox, *Non-Herculean Data: A Philosophical Intervention in a Technical Debate About Judicial Opinions as Data Sources*, 25 PROC. INT'L CONF. A.I. & L. 1, 2–3 (2025) [hereinafter Cox, *Non-Herculean Data*].

95. Cox, *Super-Dicta*, *supra* note 21, at 1618–30; Cox, *Non-Herculean Data*, *supra* note 94, at 3.

96. *See, e.g.*, Jeremy Waldron, *The Concept and the Rule of Law*, 43 GA. L. REV. 1, 52–54 (2008); John Rawls, *The Idea of Public Reason Revisited*, 64 U. CHI. L. REV. 765, 769 (1997) (explaining the ideal of public reason is realized when judges "act from and follow the idea of public reason and explain to other citizens their reasons"); *Schauer*, *supra* note 82, at 651–53 (noting relationship between explanation and justification); Cohen, *supra* note 67, at 1101–03 (arguing against sincerity qua alignment); Micah Schwartzmann, *Judicial Sincerity*, 94 VA. L. REV. 987, 1012–15 (2008) (arguing in favor of sincerity qua alignment so "their reasons [are] available for public scrutiny"); Duncan Kennedy, *Freedom and Constraint in Adjudication: A Critical Phenomenology*, 36 J. LEGAL EDUC. 518, 519 (1986); Cox, *Super-Dicta*, *supra* note 21, at 1594 n.79, 1609 nn.131–34 (collecting literature); Vredenburgh, *supra* note 60, at 222–23 (giving humans a pass because we can't *make* them give us their operative reasons); *see also* Re & Solow-Niederman, *supra* note 57, at 265–66 (arguing one problem with machine adjudicators is the inability to modulate transparency).

opinion to match, they have managed to articulate their operative reasons.[97] But that jurisprudential project is not our project today.

My point is that you *don't* get explanatory alignment from human explanations in this context, unless (ironically) a mistake has been made.[98] And so, explanatory alignment—whatever its value—does not count in favor of humans, at least in these contexts. To the extent your intuitions differ about the importance of explanatory alignment in the machine and human cases, you may wish to revisit your jurisprudential views—even if there may not be much to be done about it in the human case.[99]

I proceed as follows: First, I offer a more extended discussion of what normative uncertainty is. Second, I use a simple case to illustrate the failure of explanatory alignment as the result of normative uncertainty. Third, I return to consider how the simple case is illustrative of a broader point, that threatens Arguments from Explanation more generally.

1. What Normative Uncertainty Is

My argument begins by recognizing normative uncertainty. Recall that you can think of normative uncertainty as the kind of uncertainty that a self-driving car cannot resolve.[100] A self-driving car might be able to help resolve uncertainty about the facts and make predictions about the likely consequences of a collision based on considerations like speed and size. And grimmer—erm, purportedly "more ethical" (?!)—versions could use facial recognition to calibrate turning decisions based on the potential victims' anticipated contributions to humankind.[101] But even if your self-driving car could perfectly execute whichever rule you chose, you might still not know what the right rule is—you might not know what you should have the car do.[102] You have "normative uncertainty": uncertainty about what you ought to do.

---

97. *Cf.* Scott Altman, *Beyond Candor*, 89 MICH. L. REV. 296, 297 (1990) ("Perhaps judges should be candid but not introspective.").

98. Cox, *Super-Dicta*, *supra* note 21, at 1632–33 (recognizing that judges may act irrationally and include Super-Dicta in their opinions anyways, with bizarre consequences).

99. *See* Cox, *Super-Dicta*, *supra* note 21, at 1639 n.209.

100. *See supra* note 8 and accompanying text.

101. I'm not here to judge but, to be clear, this is a dystopian nightmare. Do you even need a cite for that? *See* Citron & Pasquale, *supra* note 9, at 4.

102. *Cf.* Aarian Marshall, *What Can the Trolley Problem Teach Self-Driving Car Engineers?*, WIRED (Oct. 24, 2018), https://www.wired.com/story/trolley-problem-teach-self-driving-car-engineers/ [https://perma.cc/Y2XH-7FKZ] (quoting autonomous vehicle developer

Normative uncertainty can come in many flavors. That is because normative uncertainty is uncertainty about the dictates of the relevant "ought," and there are many different oughts, or types of normativity.[103] The most obvious type of normativity is moral. And so, the most obvious sort of normative uncertainty, illustrated by my description of self-driving cars, is sometimes also called "moral uncertainty": normative uncertainty about the norms governing the *moral* ought.[104] And one of the easiest ways to understand, or model, such normative uncertainty is as uncertainty between competing moral theories—like uncertainty between whether consequentialism is correct (turn!) or nonconsequentialism is correct (and if so which version, since some nonconsequentialists say "turn!" and some say "don't!").[105]

But *moral* uncertainty is not the only type of normative uncertainty. For example, you might have normative uncertainty about the "doxastic" ought: about what you *ought to believe*.[106] Sometimes we are uncertain about what to believe because we don't know all the facts, a kind of descriptive or empirical uncertainty. But you might also be unsure about what to believe because you are uncertain about what you ought to believe in light of your evidence—about what you are *justified* in believing.[107] For example, you

---

Karl Iagnemma, who said "it's not clear what the right solution [to the trolley problem] is, or if a solution even exists").

103. *Cf.* JOHN BROOME, RATIONALITY THROUGH REASONING 22–54, 109–10, 116, 126–27 (2013) (distinguishing between many possible meanings of "ought").

104. *See* MACASKILL ET AL., *supra* note 4, at 1–3 (describing "moral uncertainty" as, variously, "uncertainty that stems not from uncertainty about descriptive matters, but about moral or evaluative matters" or as "referring to uncertainty about what we all-things-considered morally ought to do," as distinguished from other types of "normative uncertainty," like "uncertainty about which theory of rational choice is correct" or "about which theory of epistemology is correct"); LOCKHART, *supra* note 4, at 3 (explaining how practical deliberations should account for uncertainty about what one morally ought to do).

105. In earlier work, I explained this in terms of competing views about whether one is permitted to lie to the murderer at the door, and if so, whether one is permitted only when mere misleading won't work to save the murderer's intended victim, see Cox, *The Uncertain Judge*, *supra* note 4, at 751–54, and in terms of competing views over whether it is permissible to eat pork, and if so, whether one should only eat "happy" humanely raised pork, Cox *Non-Herculean Data*, *supra* note 94, at 2. *See also* MACASKILL ET AL., *supra* note 4, at 52 (discussing normative uncertainty about ethics of eating meat).

106. *See, e.g.*, R. Jay Wallace & Benjamin Kiesewetter, *Practical Reason*, *in* STAN. ENCYC. PHIL. 2 (Edward N. Zalta & Uri Nodelman eds., 2024) ("According to ['a different and arguably better way of understanding the contrast between practical and theoretical reason'], theoretical reflection is concerned with a normative rather than a merely descriptive question, namely with the question of what one ought, or is permitted, to believe.").

107. *Id.*

might be uncertain about the relevant epistemic norms about what evidence is relevant, about how to weigh your evidence, about what inferences you should draw, or about how assured you should be in your belief. You might also, for example, be uncertain about whether the doxastic ought can also turn on *purely* practical considerations—that is, whether you ought to believe that your boots are lucky because it is *useful* to believe that your boots are lucky.[108] Or, less radically, how practical considerations about context—about the importance of error and different types of error—should affect which inferences are permitted and what weight to give your beliefs.

Uncertainty about what you ought to believe is a type of normative uncertainty: it is uncertainty about the doxastic norms governing what you ought to believe.[109] And it is also a problem for designing self-driving cars—should small blobs be interpreted as children or cats?—even if it is not always appreciated as a "normative" one (and even if I assumed it away in my introduction of normative uncertainty!).[110]

In any event, normative uncertainty presents a practical problem: What ought you to do when you don't know what you ought to do?[111] And judges

---

108. In other words, you might be uncertain about whether the "doxastic ought" is purely an "epistemic ought," or also depends on moral and/or prudential considerations. This debate over the scope of the doxastic ought is similar to the debate over whether the jurisprudential ought can turn on moral considerations or only legal ones—and whether there is a meaningful distinction to be drawn between such considerations. For an overview of this debate with respect to the doxastic ought, see Andrew Chignell, *The Ethics of Belief*, *in* STAN. ENCYC. PHIL. (Edward N. Zalta & Uri Nodelman eds., 2018), https://plato.stanford.edu/archives/spr2018/entries/ethics-belief/ [https://perma.cc/4HA5-L9L8]. For discussion of the debate with respect to the jurisprudential ought, see, for example, SCOTT J. SHAPIRO, LEGALITY 248, 274, 280–81 (2011); Scott Hershovitz, *The End of Jurisprudence*, 124 YALE L.J. 1160, 1173–74 (2015); Mark Greenberg, *The Moral Impact Theory of Law*, 123 YALE L.J. 1288, 1306–24 (2014); *see also* Cox, *The Uncertain Judge*, *supra* note 4, at 743.

109. If what you "ought to believe" turns only on epistemic considerations, then you might call this type of normative uncertainty "epistemic uncertainty." Debates over the norms governing what you ought to believe, especially with respect to epistemic considerations, are the bread and butter of traditional philosophy.

110. One might be uncertain about whether to believe that small blob is a child or a cat because of uncertainty about purely epistemic norms—about which inference is more justifiable based on the evidence, statistics, background probabilities, etc.—or because of uncertainty about a combination of norms, some epistemic, some practical, as where one mode of interpreting the data (cat or child) will improve performance. There are difficult questions about whether to understand the latter performance issue as being about what one ought to "believe" or about practical response to uncertain beliefs. *Cf.* Jacob Ross, *Rejecting Ethical Deflationism*, 116 ETHICS 742, 744–45 (2006) (distinguishing between "belief" and "acceptance" for purposes of practical deliberation).

111. *See supra* note 4; *infra* notes 219, 222. *See generally, e.g.*, Christian Tarsney, *Moral Uncertainty for Deontologists*, 21 ETHICAL THEORY & MORAL PRAC. 505, 506 (2018) (arguing

face it too.[112] The judge's problem follows from three basic and non-cynical assumptions:

1. "Judicial decisions can be coherently criticized—that is, we speak coherently when we suggest that a judge should have decided otherwise than she did—such that we may speak of what a judge ought to do in deciding a case.
2. A conscientious judge aims to do what she ought to do in deciding a case.
3. Judges behave (or ought to behave) rationally."[113]

The first assumption is that there is such a thing as the judicial ought, such a thing as what the judge ought (judicially) to do.[114] Normative uncertainty in judicial decision-making appears when a judge is uncertain about what the judicial ought requires: she is uncertain about what she ought (judicially) to do. I do not take a position on whether what the judge ought (judicially) to do is fully determined by legal considerations, or may also depend on "moral, political, prudential, practical, or other considerations"—indeed, this may be something about which our judge is uncertain.[115]

Normative uncertainty presents a practical problem for a judge who aims to do what she ought (judicially) to do (second assumption) and who acts rationally in pursuit of that aim (third assumption): given her aim of doing

---

that "deontologists *can* say something plausible, precise, and well-motivated about decision-making under uncertainty, of both the empirical and purely moral varieties"); Andrew Sepielli, *What to Do When You Don't Know What to Do When You Don't Know What to Do . . .*, 48 NOÛS 521 (2014) (defending significance and coherence of attempts to develop theories about what to do in cases of moral uncertainty); Andrew Sepielli, *Normative Uncertainty for Non-Cognitivists*, 160 PHIL. STUD. 191 (2012) (arguing that "all otherwise plausible forms of non-cognitivism can capture something like normative uncertainty"); Ross, *supra* note 110 (arguing that deflationary ethical theories should be rejected "as a basis for guiding our actions"). *But see* Elizabeth Harman, *The Irrelevance of Moral Uncertainty*, 10 OXFORD STUD. IN METAETHICS 53, 58 (2015) (arguing that moral uncertainty is either false or uninteresting); Brian Weatherson, *Running Risks Morally*, 167 PHIL. STUD. 141, 142, 147–54 (2014) (arguing that moral recklessness is not wrong).

112. Cox, *The Uncertain Judge*, *supra* note 4, at 755–65.

113. These assumptions are taken verbatim from Cox, *Super-Dicta*, *supra* note 21, at 1587, which in turn takes them near verbatim from Cox, *The Uncertain Judge*, *supra* note 4, at 741.

114. Cox, *The Uncertain Judge*, *supra* note 4, at 741.

115. Cox, *Super-Dicta*, *supra* note 21, at 1585 ("A jurisprudence is thus not necessarily a mere interpretative theory or theory of legal reasoning—though a given jurisprudence might be."); *cf.* SHAPIRO, *supra* note 108, at 247–48 (emphasizing distinction between "legal reasoning" and "judicial decision[-]making").

what she ought (judicially) to do, what ought the judge *rationally* do when she's unsure of what she ought *judicially* to do?[116] This practical problem is the *problem* of normative uncertainty in judicial decision-making: the problem of figuring out what to do when you don't know what you ought to do.[117]

It is not my project here to defend these assumptions or the existence of the problem. I do that work elsewhere.[118] But there are a few common points of confusion that are worth elucidating.

First, is there such a thing as the "judicial ought"? Sure, there is—or at least our practices of arguing over it assume as much. Think judges should just "call balls and strikes"?[119] That's a view of the judicial ought. Think judges have broad discretion? Fine, but so long as you don't think they can wield it indiscriminately, you have a view.[120] The problem arises whether you think judges should be originalists, or textualists, or purposivists, or hermeneutic pluralists.[121] That is, so long as you think we do something more than merely shout "yay" or "boo" like fans cheering from the sidelines when judges issue opinions—that our criticisms of judges and judicial opinions have bite, whatever explains that bite—the problem gets off the ground. And if you don't think that, why care how we program our automated adjudicators—or whether we use humans or machines—other than that you simply don't "like" certain ways of proceeding?

Second, some suppose normative uncertainty only appears when there are underlying "normative" judgments to be made—when the law "runs out" and judges must turn to other factors, like morality, or when cases concern contested normative concepts.[122] But this supposition misunderstands the judicial ought. The judicial ought governs what the judge ought (judicially) do—and a view that judges ought not make "normative" judgments is a very strong view of the judicial ought!

Third, some query what normative uncertainty might look like. So, a simple illustration: Suppose that a judge believes a case should come out a certain way, but when she goes to write the opinion, it "won't write." She

---

116. Cox, *The Uncertain Judge*, *supra* note 4, at 741.

117. *Id.* at 740–41.

118. *See generally id.* (analyzing the problem of "normative uncertainty" in judicial decision-making).

119. *See id.* at 742–44; Cox, *Super-Dicta*, *supra* note 21, at 1586–87.

120. Cox, *Super-Dicta*, *supra* note 21, at 1587.

121. *See* Cox, *The Uncertain Judge*, *supra* note 4, at 743–44.

122. Cox, *Super-Dicta*, *supra* note 21, at 1587 (explaining how such a view would be mistaken); *see* Cox, *The Uncertain Judge*, *supra* note 4, at 742–44.

cannot reconcile her views on individual legal issues with what she believes is the correct outcome of the case. Perhaps, she is uncertain as between three jurisprudential approaches, which would evaluate the case as follows:

Figure 1.[123]

	Jurisprudence 1 ($p_1 = 1/3$)	Jurisprudence 2 ($p_2 = 1/3$)	Jurisprudence 3 ($p_3 = 1/3$)
Contract Issue	✓	✗	✓
Breach Issue	✓	✓	✗
Decision (Outcome Liability)	✓	✗	✗

Such situations create voting paradoxes for members of multi-member courts, and discursive dilemmas for group decision-making.[124] Indeed, this example is lifted from *The One and the Many* by Professors Lewis Kornhauser and Lawrence Sager.[125] But it seems that they might also arise for an individual judge, sitting alone, tasked with issuing a ruling—a decision and an opinion to match.[126] If she aims to be coherent, what then? What ought she (judicially) to do?

In any event, it is not my project here to defend the existence of the problem of normative uncertainty (generally, or for judges), or these three assumptions.[127] My project here is to discuss what follows from them, and how that intersects with arguments for and against a human decision. For it

---

123. Adapted from Kornhauser & Sager, *supra* note 41, at 11. The original example is framed in terms of members of a multi-member court; the table here modifies the presentation to illustrate the possibility of analogous coherence problems for a single judge.

124. *Id.*; Philip Pettit, *Deliberative Democracy and the Discursive Dilemma*, 11 PHIL. ISSUES 268, 272–73 (2001).

125. Kornhauser & Sager, *supra* note 41, at 11.

126. *See* Cox, *Super-Dicta*, *supra* note 21, at 1591–600 (explaining terms with greater detail).

127. For a defense of the problem, see Cox, *The Uncertain Judge*, *supra* note 4, at 741.

turns out that if what a judge ought (judicially) to do includes providing an explanation—providing reasons—for her decision, then her explanation will likely fail to be an aligned one between her operative reasons and the justificatory reasons she offers.[128]

2. How Normative Uncertainty Undermines Explanatory Alignment: A Simple Case

I am going to offer a simple example of how explanatory alignment might fail. The case is a little bizarre, in that it is one of complete equipoise, and so being an edge case, might be thought to be relatively insignificant as an argumentative foil. But I use the case to illustrate the point, and in the next section will turn to broader implications.

Consider a human judge deciding a case about self-driving cars, *Primo v. Turner*.[129] Primo sued Turner for injuries sustained when Turner overrode an Auto-nomous™ self-driving car's system and caused it to crash into Primo.

The facts are as follows: Turner had been sitting in a coffee shop when she saw a self-driving car careening out of control and towards a nearby intersection. The self-driving car had no passenger, but the intersection was a busy one, and a traffic jam was starting because a van full of students from the Cinque School of Law had broken down right in the center of the intersection. As a cybersecurity researcher, Turner knew how to exploit a vulnerability in Auto-nomous's software to force the car to abruptly pull over. But the exploit could only turn the car in one direction—to the right—and at that very moment, a cyclist (Primo) was in the bike lane to the right of the car. Turner realized that, if she overrode the car system, the Auto-nomous car would certainly crash into that cyclist. A card-carrying consequentialist, Turner flipped the switch and the Auto-nomous car crashed into Primo, saving the Cinque students and others who were gathering in the intersection.

Primo sued Turner for battery.[130] Turner moved to dismiss the case under the doctrines of private and public necessity, arguing that she turned the car to save the Cinque students and others in the intersection. And now, our

---

128. Cox, *Super-Dicta*, *supra* note 21, at 1581.

129. This is not—yet, or ever I hope—a real case.

130. Her attorney must have forgotten claims under the Computer Fraud and Abuse Act. 18 U.S.C. § 1030; N.Y. STATE TECH. LAW § 210 (McKinney 2025); CAL. GOV'T CODE § 8586.5 (West 2026).

dear judge, Judge Hugh Mann, must decide how to resolve the motion and on what basis. That is, Judge Hugh Mann must issue a ruling: a decision that resolves the motion, and an opinion that explains why the decision is what it is.[131]

Primo v. Turner presents exactly the kind of case about which a judge might be reasonably uncertain. It sits on the edge of private and public necessity.[132] Private necessity has provided endless fodder for tort theorists since before regular cars were invented.[133] Meanwhile, "[t]he case law on public necessity is sparse and the nature of the common law doctrine is considered 'uncertain,' 'obscure' and 'in a state of flux'"[134]—and, worse, it is *particularly* "thin and inconsistent" in cases like this, where the "claimant is treated as a means to an end or as collateral damage."[135]

Judge Mann isn't sure what to do. He'd never admit it, to be sure. But this is not a matter of pride. In fact, that he won't admit it is my point: given his aim of doing whatever it is that he ought (judicially) to do, he ought not (rationally) disclose his normative uncertainty or resolution thereof—doing so would almost certainly be self-defeating.[136]

To illustrate this, we're going to need to engage in a bit of simplification. There are far too many different ways of viewing the problem (see the vast literature[137]) and our goal is not to solve it. Similarly, there are far too many different views about how to handle such problems generally, about

---

131. Note the ambiguity about the explanation demanded. *See supra* Section I.A.

132. *See generally* Sandy Steel, *Private Right and Public Right*, 2025 CAN. J.L. & JURIS. 527 (exploring the boundaries and intersections between private right and public right).

133. Or what feels like that long. For more modern entries in the debate, see, for example, Kenneth W. Simons, *Self-Defense, Necessity, and the Duty to Compensate, in Law and Morality*, 55 SAN DIEGO L. REV. 357, 363–64 (2018); Laura A. Heymann, *Trolley Problems, Private Necessity, and the Duty to Rescue*, 60 SAN DIEGO L. REV. 1, 16–27 (2023); Sandy Steel, *Saving* Private Wrongs, 14 JERUSALEM REV. LEGAL STUD. 1, 8 (2016); ARTHUR RIPSTEIN, PRIVATE WRONGS 146–55 (2016).

134. Samuel Beswick, *The Defense of Public Necessity,* 88 MOD. L. REV. 973, 975–76 (2025) (footnotes omitted) (quoting Graham Virgo, *Justifying Necessity as a Defence in Tort Law*, *in* DEFENCES IN TORT 212, 212 (Andrew Dyson et al. eds., 2015); John P. Finan & John Ritson, *Tortious Necessity; The Privileged Defense*, 26 AKRON L. REV 1, 8 (1992); JAMES GOUDKAMP, TORT LAW DEFENCES 182 (2013)); Illert v. N. Adelaide Loc. Health Network Inc. (Modbury Hosp.) [2016] SASC 186; Rigby v. Chief Constable of Northamptonshire [1985] 1 WLR 1242, 1254; Binsaris v. N. Territory [2020] HCA 22, [43]–[45]; JOHN MURPHY, STREET ON TORTS 306 (12th ed. 2007); Raissi v. Metropolitan Police Comm'r [2007] EWHC (QB) 2842 [39]–[44]; *Re A (Children)* [2001] Fam 147 at 220–22).

135. Beswick, *supra* note 134, at 976.

136. *See* Cox, *Super-Dicta*, *supra* note 21, at 1616–19.

137. Simons, *supra* note 133, at 377; Heymann, *supra* note 133, at 16; *see also* Steel, *supra* note 133, at 8.

whether there is discretion and how to exercise it, and so forth.[138] And in reality, there are far too many options available to the judge about how to resolve the motion (e.g., grant, deny, delay or avoid by referring to mediation).

I'm going to make two assumptions.

First, I'm going to model Judge Mann's uncertainty as between two views about what he ought (judicially) to do—two "jurisprudences."[139] A "jurisprudence" is "an all-things-considered theory about what a judge ought (judicially) to do."[140] "An all-things-considered theory means just that: a jurisprudence is a theory about what a judge ought (judicially) to do in light of all relevant legal, moral, political, prudential, practical, or other considerations, whatever those may be."[141] "A jurisprudence is thus not necessarily a mere interpretative theory or theory of legal reasoning—though a given jurisprudence might be."[142] And a jurisprudence could be multi-ordered: it may be that what "a judge ought (judicially) do is to follow some first-order procedure, and if that doesn't resolve what they ought to do, then they ought (judicially) follow some second-order procedure"—e.g., look to purpose—"and if *that* doesn't resolve what they ought [judicially] to do, then they ought (judicially) follow some third-order [decision] procedure"—like appealing to morality—"and so forth."[143]

Second, I'm going to assume that there are a limited number of rulings he is deciding between:

> A. Decide in favor of Primo *for the reason that*, while the case implicates public necessity, public necessity does not bar recovery for damages where, as here, a plaintiff is used as mere means to save others.[144]

---

138. *See* RONALD DWORKIN, LAW'S EMPIRE 225–75 (1986); JOSEPH RAZ, THE AUTHORITY OF LAW 37–52 (1979). *See generally* H.L.A. HART, THE CONCEPT OF LAW (3d ed. 2012); Ronald M. Dworkin, *The Model of Rule*s, 35 U. CHI. L. REV. 14 (1967); Ronald Dworkin, *Hard Cases*, 88 HARV. L. REV. 1057 (1975); Scott J. Shapiro, *On Hart's Way Out*, 4 LEGAL THEORY 469 (1998); Scott J. Shapiro, *The "Hart-Dworkin" Debate: A Short Guide for the Perplexed*, *in* RONALD DWORKIN 22 (Arthur Ripstein ed., 2007).

139. Cox, *The Uncertain Judge*, *supra* note 4, at 742–44.

140. Cox, *Super-Dicta*, *supra* note 21, at 1585.

141. *Id.*

142. *Id.*; *cf.* SHAPIRO, *supra* note 108, at 247–48 (distinguishing "legal reasoning" from "judicial decision[-]making").

143. Cox, *Super-Dicta*, *supra* note 21, at 1645.

144. *See* Beswick, *supra* note 134, at 986 (arguing in favor of this view).

B. Decide in favor of Turner *for the reason that* the case implicates public necessity, and public necessity is a complete defense.[145]

Jurisprudence 1 reflects the view of Professor Samuel Beswick, and supports Ruling A. Jurisprudence 2 reflects the conventional view of the doctrine of public necessity, and supports finding Ruling B.[146]

I am also going to assume, for now, that Jurisprudence 1 and Jurisprudence 2 agree about the stakes.[147] That is, while the jurisprudences diverge in their reasoning and so decision, they agree that it would be an obvious and significant wrong to decide for the wrong party, or for the wrong reasons. This divergence is not implausible. After all, Jurisprudence 1 favors Ruling A, based on the principle that rule of law requires preventing people—their literal bodies—from being used as mere means, a principle with solid Kantian bona fides.[148] Meanwhile, Jurisprudence 2 views the defense as complete under black letter law, such that Ruling A is not only incorrect but would severely distort the law by creating an exception where there is none and where the underlying legal materials are plain (even if some rogue commonwealth courts previously got it wrong!).[149]

We can illustrate Judge Mann's choice using a decision matrix. His credences—assessments that one or the other Jurisprudence gives the correct account of what he ought (judicially) to do—are represented as probabilities (p_i).[150] For ease of illustration, I'm going to assume that Judge Mann is equally confident in each of these jurisprudences ($p_1 = p_2 = .5$).

---

145. *See id.* at 981–87 (noting that the consensus view of public necessity is that it serves as a "complete" privilege).

146. *See id.*

147. I will return to this assumption in Section II.B. The same is true for more lop-sided stakes and credences but would unnecessarily complicate the point about explanatory alignment.

148. *Cf.* Beswick, *supra* note 134, at 980–89 (suggesting necessity may provide a defense where the interference with another might be "impliedly consent[ed] to"); IMMANUEL KANT, GROUNDWORK OF THE METAPHYSICS OF MORALS 4:429, in IMMANUEL KANT, PRACTICAL PHILOSOPHY 41, 80 (Mary Gregor ed. & trans. 1996).

149. *See* Beswick, *supra* note 134, at 974–76.

150. For more about credences, see Cox, *Super-Dicta*, *supra* note 21, at 1589; Cox, *The Uncertain Judge*, *supra* note 4, at 779 n.141.

Figure 2. *Primo v. Turner*: Scenario I

	Jurisprudence 1 ($p_1 = .5$)	**Jurisprudence 2** ($p_2 = .5$)
Ruling A	Right	Obviously wrong
Ruling B	Obviously wrong	Right

What ought Judge Mann (rationally) do given his uncertainty about what he ought (judicially) to do? There is an easy option in cases like this.

First, note that the case—from Judge Mann's point of view—is in complete equipoise: his credences are split and the jurisprudences agree about the stakes. From where he sits, he believes he has an equal chance of getting it right or obviously wrong either way. And what is rational to do when options are in complete equipoise?

You can flip a coin.[151]

Suppose Judge Mann, acting rationally, flips the coin and it lands tails—issue Ruling B. Should Judge Mann disclose that in his opinion? Should he explain his operative reasons—that he was uncertain about how to decide a case like this, assessed the options to be in complete equipoise, and then flipped a coin?

If by now you are thinking "obviously no!"—shock, horror—that is probably a disproportionate reaction, but you're not wrong about the disclosure. Here's the (more measured) reaction: disclosing the coin flip would be self-defeating.

To see this, think about what such a ruling might look like. Here's an example:

> B*. Decide in favor of Turner *for the reason that* the case implicates public necessity, and public necessity is a complete defense, and the reason I so conclude is that the case is actually hard and not obvious and so I flipped a coin.

---

151. *See* Adam M. Samaha, *On Law's Tiebreakers*, 77 U. CHI. L. REV. 1661, 1690 n.79 (2010); James D. Nelson & Micah Schwartzman, *Second-Order Decisions in Rights Conflicts*, 109 VA. L. REV. 1095, 1134 (2023).

The first thing to note about this ruling—other than its strangeness—is that it is a distinct ruling. That is, this coin flip ruling is *not* Ruling B (which gives the first set of reasons) but a separate ruling, call it Ruling B*. In other words, Ruling B* is a different option within the choice set:

Figure 3. *Primo v. Turner*: Scenario I (with Ruling B*)

	Jurisprudence 1 ($p_1 = .5$)	**Jurisprudence 2** ($p_2 = .5$)
Ruling A	Right	Obviously wrong
Ruling B	Obviously wrong	Right
Ruling B*	Obviously wrong	Obviously wrong

So, should Judge Mann issue Ruling B*? That is, given that Judge Mann does not know what he ought (judicially) to do, ought he (rationally) to issue Ruling B*? The answer is no. Ruling B* is transparent, sure. But while Judge Mann is uncertain about what he ought (judicially) to do, he can at least eliminate wrong options. And Ruling B* is obviously wrong.

Why is this? Recall that Judge Mann's options only *appeared* (to him) to be in complete equipoise. He did not believe that the case was actually in equipoise. Rather, he knows that the stakes are high and the answer would be "obvious" if only he weren't uncertain about which theory—which jurisprudence—provides the correct account of what he ought (judicially) to do. But the one thing he is sure of—the one thing on which the jurisprudences agree—is that this case is *not a coin toss*. And so, even if Judge Mann resolves his uncertainty by flipping a coin—as is rational to do—he still should not issue a ruling saying that that is what he did. Doing so would be irrational, because it is self-defeating.[152]

---

152. Why not say the case reveals that some third jurisprudence, Jurisprudence 3, is correct, which says that Ruling B* is correct? Or some other solution? Sure, you can fight the hypothetical—indeed, the phenomenology of many cases may be just like this since I doubt anyone has a complete jurisprudence. But jurisprudences are not infinitely malleable. And so, you will encounter a case where there is no Jurisprudence 3. For discussion of malleability, see Cox, *Uncertain Judge*, *supra* note 4, at 67–68. For discussion of why a Jurisprudence 3 as

What does this mean? Judge Mann flips a coin, and issues Ruling B. But Ruling B does not explain the Judge's operative reasons for the decision—based on what features and by what logic he actually decided for the defendant. Ruling B gives different reasons. Judge Mann's explanation—his opinion—does not exhibit alignment.

3. From the Simple Case to a Broader Difficulty

This example may seem fanciful. The thought of a judge flipping a coin to decide a difficult case might seem laughable.[153] And a case that is in complete equipoise might seem unusual. Allow me to take each in turn.

Consider what might seem like a more familiar approach, one that has been proposed in the constitutional law literature on hard cases—both as a normative and descriptive matter.[154] For example, suppose Judge Mann considered the ability of each party to cover the damage, and used that to break the tie.[155]

Would he say as much? No, not if he is acting rationally. For it would still be self-defeating to issue a ruling that said as much. For one, it would say the case was hard when that is the one thing Judge Mann knows to be untrue. But it would also be obviously wrong in other ways: such a ruling would be obviously wrong by the lights of Jurisprudence 1, because it suggests that people's bodies can be used as mere means provided they have good health insurance. And it would be obviously wrong by the lights of Jurisprudence 2, because, like Ruling A, it totally distorts the very straightforward doctrine of public necessity.

This case might also seem fanciful because complete equipoise is rare. But the same holds true for more complicated cases—for more lop-sided credences—and more sophisticated ways of resolving normative uncertainty, like using expected value theory or employing various harm-avoidance approaches proposed in the constitutional law literature on hard cases.[156] For example, I've previously offered a decision-theoretic

---

posited—that would hedge by issuing Ruling B*—is almost certainly a nonstarter, see *id.* at 771–75.

153. *See* Samaha, *supra* note 151, at 1690–91 (noting this reaction but debunking it).

154. *See* Aaron Tang, *Harm-Avoider Constitutionalism*, 109 CALIF. L. REV. 1847, 1849–50 (2021); Cox, *Super-Dicta*, *supra* note 21, at 1613–14.

155. *See* Tang, *supra* note 154, at 1849 (describing "harm-avoider constitutionalism").

156. *See* Cox, *Super-Dicta*, *supra* note 21, at 1607–16 (discussing the merits and demerits of two possible approaches for addressing normative uncertainty in judicial decision-making, the "Maximization Approach" and "Conflict Avoidance"); *see also* Charles L. Barzun &

reconstruction of *Brown v. Board of Education*[157] that is consistent with the unanimity of the court despite divergent views among the justices about the judicial ought.[158] And I showed how a similar analysis applied to *Google v. Oracle*[159] revealed it would frequently be self-defeating to reveal such uncertainty and its resolution.[160]

I also do not mean to suggest that judges always refrain from noting their normative uncertainty. Perhaps, you know it when you see it.[161] But such disclosure comes at a cost.[162] And then, ironically, Arguments from Explanation only favor the humans when the human is a schlub.[163]

There might be a final worry, that these cases reflect truly difficult ones, of a sort where machines are wildly inappropriate.[164] And that the real disputes over use of machines occur in easy cases. But neither should give the humanist comfort.

For one, if this is only a problem in hard cases, Arguments from Explanation would seem to fail precisely where they seem most important: those hard cases in which a machine can give a plausible-sounding opinion that passes what Professor Eugene Volokh has called the "Ordinary Schlub Criterion."[165] Before recent advances in AI, we perhaps didn't need to worry about these hard cases because machines were not sufficiently sophisticated to "write" an opinion.[166] But now, we'll need a better story about why the disconnect between the machine-generated "opinion" and the machine's

---

Michael D. Gilbert, *Conflict Avoidance in Constitutional Law*, 107 VA. L. REV. 1, 3 (2021); Tang, *supra* note 154, at 1849–50.

157. 347 U.S. 483 (1954).

158. Cox, *The Uncertain Judge*, *supra* note 4, at 788–96.

159. 593 U.S. 1 (2021).

160. *See* Cox, *Super-Dicta*, *supra* note 21, at 1625–34 (discussing why it would be self-defeating to disclose uncertainty and reasoning on an expected value approach); *id.* at 1650–51 (discussing why it would be self-defeating to disclose use of a harm-avoidance type approach to resolve normative uncertainty).

161. *See id.* at 1631–41 (discussing Jacobellis v. Ohio, 378 U.S. 184, 197 (1964) (Stewart, J., concurring)).

162. *Id.* (discussing the transparency-objectivity trade-off).

163. *See* Paul Gerwitz, *On "I Know It When I See It"*, 105 YALE L.J. 1023, 1024, 1025 n.7 (1996) (noting Justice Stewart came to regret his statement in *Jacobellis* that "I know it when I see it, and the motion picture involved in this case is not that").

164. *Cf.* Huq, *Human Decision*, *supra* note 2, at 619 (holding that machines should not be used as decision makers where the decision "entails ethical or otherwise morally charged judgments").

165. Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1139 (2019).

166. *See infra* Part II.

operative reasons does not matter—for there is a similar disconnect between the human-generated opinion and the human's operative reasons.

How interesting: even if you take the point to be limited by the complexity of my examples, the result is that Arguments from Explanation fail precisely when you might need them most.

But even if the real dispute over machine use remains the easy cases to which we know the answer, there is another lingering concern: the judge only does not recognize her normative uncertainty because all the jurisprudences in which she has some level of credence agree. And so, the human opinion does not reveal whether the judge's reasoning is deeply aligned, whether the judge follows the correct jurisprudence or an illicit one that got it right in this case.

In machine terms, this is one aspect of the problem of value alignment: the machine's decisions are consistent with multiple objective functions, some of which we would endorse and others which we wouldn't. But we don't know which the machine is acting on. And we want an aligned explanation because we are worried about the cases where the machine will go wrong—possibly terribly so—because its objective function, its operative reasons, are not what we took them to be. An aligned explanation would alleviate this problem, but an aligned explanation is unavailable for certain technologies. And so, Arguments from Explanation would seem to favor humans, even in such easy cases.

But my point is that we do not have this in the human case, with similar results. As with machines, we won't discover the lack of "value alignment" until later. Why would Arguments from Explanation favor humans even in these easy cases, if humans also do not provide the depth of explanatory alignment demanded of machines?

In any event, I have used a simple case because my goal here is not to defend the lack of explanatory alignment in the human case—I do that work elsewhere[167]—but to show how that result intersects with the "right to a human decision" debate. And the simple case shows the point: when a judge is uncertain about what they ought (judicially) to do, and they act rationally in light of that uncertainty, their actual operative reasons for deciding the case and the stated reasons often diverge. That is, in the case of a *conscientious human judge* acting rationally under conditions of uncertainty, there is likely a lack of alignment between the judge's operative reasoning and the explanation of that reasoning as expressed in the opinion.

---

167. *See* Cox, *Super-Dicta*, *supra* note 21, at 1600.

This lack of alignment in the human case undermines the Argument from Explanation in this context. Whatever the merits of a call for an aligned explanation from a machine, the inability of a machine to provide one does not provide a reason to favor a human in these contexts. The failure of the argument does not result from cynical premises about the human case. It does not depend on the "Ordinary Schlub Criterion," that the relevant comparator is not "the best of the best," but "the average person whom we are considering replacing."[168] It does not depend on myriad other considerations that have been raised in the literature and called into doubt by techno-optimists, through references to the "black box" of the human brain and certain forms of jurisprudential realism.[169] The lack of alignment appears in even the *ideal* human case.

To be sure, the absence of aligned explanations may bear on the importance of judicial character in selecting a decisionmaker. But that is a different sort of argument.[170] The Argument from Explanation maintains that machines should not replace humans where explanations are owed because only humans can provide the requisite sort of explanation. And so, the lack of alignment in the even the *ideal* human case provides a powerful reason to doubt that the Argument from Explanation favors humans in contexts where explanations are owed for decisions made.[171]

II. A BETTER DECISION?

Arguments from Explanation had appeared to provide humanists with a strong normative ground against the use of machines, especially in contexts like adjudication where explanations are owed. And Arguments from Explanation appeared to grow increasingly forceful as new machine learning technology grew increasingly inscrutable and nonintuitive: aligned

---

168. Volokh, *supra* note 165, at 1139.

169. Coglianese, *supra* note 2.

170. One might instead argue that the absence of aligned explanations increases the "importance of judicial character" in selecting a decisionmaker. Cox, *Super-Dicta*, *supra* note 21, at 1584 n.35. The significance of character may, in turn, inform the selection of human or machine—at least, if we are better able to evaluate the dispositions of one over the other. But that is a different sort of argument than Arguments from Explanation. With thanks to Felix Wu for discussion.

171. Does this mean the lack of alignment reveals a problem about the human case? I'm not sure that that is the right question to ask. But to the extent you feel unease in the machine case, it seems you should feel unease in the human case, too, for all the reasons that alignment is desirable in the machine case. Alas, just because something is desirable doesn't mean you can get it.

explanations had become *necessarily* unavailable for machine decisions. But if humans cannot provide an aligned explanation when they respond, rationally, to normative uncertainty in contexts like adjudication, then the humanist's Argument from Explanation no longer favors human decisionmakers over machine ones, at least not in its current form.[172]

The failure of Arguments from Explanation might seem particularly disappointing in the face of the machinist's Better Decision Argument.[173] The machinist's Better Decision Argument reduces the humanists' putative normative objections—objections that there are normative limits to machine use based on principles like equity, dignity, and autonomy, or on institutional effects—to technical ones.[174] The machinists do so by arguing that humanists' principles do not ground a right to a human decision, but a right to a "better" decision—to the better of a human decision or a "well-calibrated" machine decision.[175] That is, the machinist's Better Decision Argument claims that all the humanists have managed to show is that the machine is not sufficiently well calibrated—and whether the machine *can* be so calibrated turns on technical, not normative, limitations.[176]

The humanist Argument from Explanation had seemed increasingly strong because it turned on a technical limitation: machines necessarily cannot provide an aligned explanation.[177] But as just argued, humans cannot (rationally) provide an aligned explanation either—at least not in contexts where explanations for decisions are owed, and so where Arguments from Explanation might be thought to have most force.[178]

And so, having defanged what seemed like one of the humanist's strongest arguments against replacing humans with machines in contexts, like adjudication, where an explanation is owed, the question arises: What could ground such an argument now?

To the extent the Better Decision Argument is successful against other humanist principles, it would seem the right to a human decision is dead.[179]

---

172. *Supra* Section I.B.

173. *See* Huq, *Human Decision*, *supra* note 2, at 619.

174. *See, e.g.*, *id.*; Coglianese, *supra* note 2.

175. *See* Huq, *Human Decision*, *supra* note 2, at 619; Coglianese, *supra* note 2.

176. *See* Huq, *Human Decision*, *supra* note 2, at 619; Coglianese, *supra* note 2.

177. *Supra* Section I.A.

178. *Supra* Section I.B.

179. Or rather, the right's normative basis is dead. The right might not be legally dead. *See* Regulation 2024/1689, 2024 O.J. (L 1689) 1, 18 (EU); *id.* at 1 (identifying source of law).

The only limit is technical, not normative.[180] Normative uncertainty gives machinists the advantage, or so it seems.

But what normative uncertainty gives with one hand, it takes with another. Normative uncertainty may have undermined humanist Arguments from Explanation. But the problem of normative uncertainty also undermines the machinists' Better Decision Arguments in a crucial way: asserting the right to a "better" decision is cold comfort when we have uncertainty about what a better decision is.

In Section A, I will expand on the machinist's Better Decision Argument. I will also explain why a "normative judgment" caveat that some quasi-machinists have asserted also seems newly vulnerable to the Better Decision Argument. In Section B, I will explain why Better Decision Arguments fail.

A. The Right to a Better Decision

The Better Decision Argument generally begins by observing that humans are not—yet—totally out of the loop. Rather, human judgments—human decisions—are baked into the machine.

Return to the trolley problem and self-driving cars. To (re)state the obvious, self-driving cars face real-life trolley problems. They must be programmed to avoid collisions with pedestrians and other obstacles. Avoiding them is easy if nothing else is around. It becomes more complicated, both technically and ethically, if a collision is unavoidable—a trolley problem.

Self-driving cars do not solve trolley problems *ab initio*. Rather, we—the human developers and regulators—must make these normative choices, or cede them to chance. As Professor Huq explains:

> The dearth of reasoned judgments in machine decisions . . . is something of an optical illusion. It is not so much that such judgments are wanting. Rather, they have already been embedded

---

180. Huq, *Human Decision*, *supra* note 2, at 611–12 ("Instead of being derived from normative first principles, limits to machine decision making are appropriately found in the technical constraints on predictive instruments."). As discussed *infra*, it is not clear that Professor Huq's "no normative judgment" exception to the Better Decision Argument continues to apply in the age of LLMs capable of imitating moral reasoning. And, as discussed *supra* note 36, technical limitations may include institutional limitations—institutional design constraints that affect how well institutions may govern deployment of machine decisionmakers. *See* Huq, *Constitutional Rights*, *supra* note 8, at 1907–12.

> into a classifier by the time that an algorithm is working in the world.[181]

Because human judgment will still be required at some point in the process, the real questions are who, when, and how.[182] What the humanists really want is not just human involvement, but a particular type of human decision: "human judgment *at a particular moment*"—namely, "after a machine-learning instrument in the wild encounters and classifies a human actor."[183] For ease of exposition, I will call this particular moment the "time of encounter."[184]

Once it is understood that the real debate is not over *whether* a human should be involved, but *when* and *how*, the argument goes, the right to a "human" decision dissolves into arguments about whether machines can provide a "better" decision at the time of encounter.[185] That is, the basic

---

181. Huq, *Human Decision*, *supra* note 2, at 646, 651, 675; *see also* Crootof et al., *supra* note 25, at 434, 440–45 ("It's humans all the way down. But these and other myriad influential forms of human involvement fade into the unregulated background when we focus overmuch on a discrete application of a system or its particular results."). *See generally* Lehr & Ohm, *supra* note 2 (arguing for greater focus on choices made during development and fine-tuning of machine-learning models).

182. *See, e.g.*, Huq, *Human Decision*, *supra* note 2, at 686 (suggesting that machine learning use warrants "a modicum of sensitivity to the capabilities and limits of machine learning"); Kaminski & Urban, *supra* note 2, at 1964 (assessing right to contest AI decisions); Green, *supra* note 89, at 11–12 (suggesting that government algorithms should be regulated via institutional rather than human oversight); Jon Kleinberg et al., *Human Decisions and Machine Predictions* 24–25 (Nat'l Bureau of Econ. Rsch., Working Paper No. 23180, 2017) (explaining that AI may assist human judges in better assessing release outcomes).

183. Huq, *Human Decision*, *supra* note 2, at 651 (emphasis added). I take this timing to include the right to contest a machine decision to a human decisionmaker for something more than just ensuring that the machine performed as intended. *See, e.g.*, Kaminski & Urban, *supra* note 2, at 1986. There is, after all, a big difference between de novo review and merely confirming the machine's validation metrics. *See id.* ("Aziz Huq has gone further, arguing that due process rights should not apply even to state action—or rather, that instead of individualized challenges to algorithmic decision-making, individuals should be able to challenge whether the algorithm is systemically 'well-calibrated.'"). *But see* Huq, *Constitutional Rights*, *supra* note 8, at 1950 (suggesting that much about due process in this context remains to be worked out).

184. I use "time of encounter" instead of "time of classification" to avoid ambiguities in what might be meant by "classification." There are difficult ambiguities about what "time of encounter" really means, that relate to difficulties in defining "scope of choice," but I hope my meaning in context is sufficiently clear for sake of abstract argument. *See infra* Section III.A; *see also* Lehr & Ohm, *supra* note 2, at 655 (calling this "the running model"); *cf.* Crootof et al., *supra* note 25, at 433–34, 440–45 (critiquing definition of "human in the loop" as "an individual involved in a single, particular algorithmic decision").

185. The basis for comparison is not always clear. For example, are machines "better" if they are better on average? Consistently? In important cases? And so on.

strategy of the Better Decision Argument is to debunk the claim that what matters is a *human* decision at the time of deployment by arguing roughly as follows: a number of principles are thought to ground the right to a human at the time of encounter. But on closer inspection, what they amount to is a demand for a decision that meets certain criteria or respects certain values. And whether machines can provide this depends on technical limitations, not normative ones.

Take bias. Self-driving cars depend on sensors and sensors are notoriously biased.[186] But if self-driving cars are actually biased in this way, the appropriate solution is not necessarily to switch to human drivers. Humans are often biased too! Rather, Huq and others claim, the solution is to fix the machine.[187] The question then will become whether the machine *can* adequately address these inequities—or at least, whether the machine can address them better than humans in a given context. But that is about technical limitations, not normative ones. And to the extent a machine cannot satisfy multiple measures of equal treatment, this limitation on machines is *also* a limitation for humans.[188] In short, arguments based on bias do not ground a right to a human but to the *better* of a human or a "well-calibrated machine."[189]

To offer another example, consider dignity or autonomy. These values are frequently cited as grounding a right to a human decision. But as Professor Huq and others have observed, a human decisionmaker is not always dignity- or autonomy-enhancing.[190] This much may be obvious in the case of self-driving cars. But it is also the case in situations where dignity and autonomy have greater pull, like benefits determinations and

---

186. *See, e.g.*, Zhan & Wan, *supra* note 7, at 6.

187. *See* Huq, *Human Decision*, *supra* note 2, at 631, 650–51; Coglianese, *supra* note 2.

188. *See* Jon Kleinberg et al., Inherent Trade-Offs in the Fair Determination of Risk Scores 17 (Nov. 17, 2016) (unpublished manuscript) (on file with arXiv), http://arxiv.org/abs/1609.05807 [https://perma.cc/U5RH-R3XP] ("[E]xcept in highly constrained special cases, it is not possible to satisfy these three [fairness] constraints simultaneously; and moreover, a version of this fact holds in an approximate sense as well."); Sam Corbett-Davies et al., Algorithmic Decision Making and the Cost of Fairness 1 (June 10, 2017) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/1701.08230 [https://perma.cc/JJ35-Q4SW]; *cf.* RUSSELL & NORVIG, *supra* note 32, at 983 ("No entity—human or machine—can prove things that are impossible to prove.").

189. Huq, *Human Decision*, *supra* note 2, at 686; Coglianese, *supra* note 2.

190. *See* Huq, *Human Decision*, *supra* note 2, at 653; *see also* Leslie Meltzer Henry, *The Jurisprudence of Dignity*, 160 U. PA. L. REV. 169, 189–90 (2011); Vicki C. Jackson, *Constitutional Dialogue and Human Dignity: States and Transnational Constitutional Discourse*, 65 MONT. L. REV 15, 37–38 (2004); Richard H. Fallon, Jr., *Two Senses of Autonomy*, 46 STAN. L. REV. 875, 877–78 (1994).

dispute resolution, where the affected person is the subject of the decision, not merely along for the ride. A human administrator can treat a person like a crunched number as easily as a machine can.[191] Machines may not smile, but they also don't judge.[192] And while quick human review may be cursory, quick machine review might not be: a machine can crunch numbers faster—with greater accuracy, accounting for more features—than a human. And sometimes speed is autonomy- and dignity-enhancing, even if it comes at the cost of some accuracy or the "human touch."[193]

I won't canvas all the arguments advanced in favor of humans, or assess the success with which Huq, Coglianese, and others have refuted them.[194] My point is to sketch the argumentative strategy at which I take aim: the argument that some standard list of principles and problems—bias, autonomy, dignity, institutional effects—support not a *human* decision but a *better* decision. Call machinist arguments of this sort "Better Decision Arguments."

Of course, machines do require deciding in advance and there are tradeoffs to doing so. But the downsides have less to do with the timing of the normative judgment than the lock-in.[195] Humans also make normative decisions in advance; they need to, if they have any prayer of acting on them.[196] It's called preparation. One of the principal advantages of trolleyology is to work out scenarios in advance for planning purposes: to plan to avoid them, and to plan how one ought to respond to them.[197] Humans also work out rules of thumb for driving—algorithms—like Brake for Moose (even if you'll get rear-ended) and the Squirrel Rule (don't swerve for squirrels).

Wait, what? Where did the squirrels come from? Exactly. That is the point. When you're driving, things pop out at you. If you swerve for

---

191. Huq, *Human Decision*, *supra* note 2, at 659–60.

192. *See id.*

193. *Cf.* Kamm, *supra* note 7, at 89 (arguing that more lives may be saved by using self-driving cars than by ensuring they respond appropriately to the trolley problem). *See generally* Yonathan Arbel & David A. Hoffman, *Generative Interpretation*, 99 N.Y.U. L. REV. 451 (2024) (arguing that LLMs may provide greater contextual analysis of a parties' bargain without the inefficiencies of traditional methods).

194. *See* Solove & Matsumi, *supra* note 11, at 1938–39; Coglianese, *supra* note 2.

195. *See* Solove & Matsumi, *supra* note 11, at 1928.

196. *See* Samuel Gorovitz, *Bioethics and Social Responsibility*, 60 MONIST 3, 14 (1977) ("Decisions in clinical medicine do not await philosophical reflection.").

197. *See* Solove & Matsumi, *supra* note 11, at 1926; *see also* Kamm, *supra* note 7, at 89; Kaminski & Urban, *supra* note 2, at 1972.

squirrels, you'll end up in a ditch, and probably still hit them anyways.[198] But your instinct will be to swerve. So, it is helpful to have thought of this in advance. This is one of the reasons trolleyology is sometimes used to train new drivers.[199] And one reason for favoring rules over standards.

The trouble with deciding in advance is that anticipating novel events is difficult. Developers may not think of a trolley scenario when programming a rule or training a machine.[200] Accordingly, algorithms may miss relevant features, either because they are missing from programmed rules, or from the data set. To the extent the street encounter presents something new—does the Squirrel Rule work for kangaroos?[201]—this will be a cost of using an algorithm rather than a human, to the extent a well-trained human can cope with kangaroos on the fly.[202] Modern machines can apply standards, but how well?

But Professor Huq's point is that this may well be the right trade-off to make.[203] Even if a human driver were able to cope with kangaroos better than the machine, this fact alone does not yet resolve whether to use a human.[204] Kangaroos may be sufficiently low probability events that it is better, on the whole, to use a kangaroo-blind machine than to count on a

198. Squirrels are not stationary. *See* Bruce Cox School of Driving.

199. Or at least, this driver. In "The Points System of Driving," you earn points by identifying increasingly outrageous potential collisions and ways of hitting or avoiding them based on the desert of those involved: "1 point for the stroller, 10 points if you avoid the stroller and nail the thoughtless parent pushing it out into the sidewalk to look both ways." Sick, I know. The assessment of parental desert was lacking. But the point was to identify—and so learn to anticipate—choices made by others that could lead to accidents.

200. Crootof et al., *supra* note 25, at 463. The difficulty of extrapolating from a limited data set to new scenarios is also one reason scholars have advanced for the poor fit of machines to common law adjudication. *See* Fagan & Levmore, *supra* note 25, at 11–14.

201. No, it turns out. Kangaroos are dangerous. Liz Ginis, *Why Roos Attack and What You Should Do to Avoid It*, AUSTL. GEOGRAPHIC, (Sept. 15, 2022), https://www.australiangeographic.com.au/nature-wildlife/2022/09/why-roos-attack-and-what-you-should-do-to-avoid-it/?passthrough=true [https://perma.cc/4SX5-PCL8]..

202. Where do kangaroos come from?! Actually, the AI literature. *See, e.g.*, Kaminski & Urban, *supra* note 2, at 1972 n.83 (citing Evan Ackerman, *Autonomous Vehicles vs. Kangaroos: The Long Furry Tail of Unlikely Events*, IEEE SPECTRUM (July 5, 2017), https://spectrum.ieee.org/cars-that-think/transportation/self-driving/autonomous-cars-vs-kangaroos-the-long-furry-tail-of-unlikely-events/ [https://perma.cc/S27X-7RBT] (explaining how autonomous vehicles struggle with kangaroos)).

203. *See* Huq, *Human Decision*, *supra* note 2, at 686.

204. *See* Kaminski & Urban, *supra* note 2, at 1972–73.

human to follow the Squirrel Rule in the much more common case of squirrels.[205] Professor Frances Kamm, master of trolleyology, agrees.[206]

Again, I won't here canvas the principles thought to ground a human decision, or the Better Decision Arguments against. My aim is to simply sketch, in broad strokes, the argumentative strategy.

There is one important caveat to Better Decision Arguments, however. And that is a concession that machines might still be an inappropriate tool for certain contexts, like adjudication, that involve "the presence of normative shades in many [of the] matters" to be "resolved."[207]

This caveat is sometimes made by fiat.[208] But technical considerations linger in its background, as it was too difficult to automate (or imitate) such complex reasoning until quite recently:

> [M]any decisions related to the law fall outside the domain of the technologically plausible. . . . Machines cannot (yet?) resolve the difficult and inescapably normative questions of aggregation, distribution, and belonging that characterize politics.[209]

But technical capabilities have advanced. What was once the stuff of "fanciful speculation"—machines capable of writing basic ethics discussions or superficially plausible opinions in response to natural language briefs—are on the scene.[210] What is more, judges have begun to

---

205. Which turns out to be quite hard. *See* Huq, *Human Decision*, *supra* note 2, at 634–36 (discussing the difficulty of not reacting to obstacles).

206. *See* Kamm, *supra* note 7, at 89 (noting the lives that can be saved by self-driving cars).

207. Huq, *Human Decision*, *supra* note 2, at 686.

208. *See id.*

209. *Id.* Professor Huq wrote this in 2020 before the massive leap forward in generative AI.

210. *See* Adam Unikowsky, *In AI We Trust*, ADAM'S LEGAL NEWSL. (June 8, 2024) [hereinafter Unikowsky, *Trust*], https://adamunikowsky.substack.com/p/in-ai-we-trust [https://perma.cc/HU4C-JMT7] ("AI is already able to decide cases correctly."); Adam Unikowsky, *In AI We Trust, Part II*, ADAM'S LEGAL NEWSL. (June 16, 2024) [hereinafter Unikowsky, *Trust II*], https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii [https://perma.cc/8R66-F4TN] ("The results were otherworldly. Claude is fully capable of acting as a Supreme Court Justice right now."). *See generally* Arbel & Hoffman, *supra* note 193 (proposing using LLMs to interpret contractual meaning); *see generally also, e.g.*, Brian Leiter, *How Much Trouble Are We in with the ChatGPT5?*, LEITER REPS.: A PHIL. BLOG (Sept. 21, 2025), https://leiterreports.com/2025/09/21/how-much-trouble-are-we-in-with-the-chatgpt5/ [https://perma.cc/W2G7-KJ2V] (sharing example essay produced in response to prompt shared by philosophy professor Jason Bridges).

While legal AI is superficially impressive, problems like hallucination and validation present significant challenges. *See* Neel Guha et al., LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models 4, 32, 104 (Aug. 20, 2023) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2308.11462

experiment with them.[211] And researchers work to pair new techniques with old, in hybrid models, that might leverage the power of the two modes.[212]

Given these technical advances, it is not clear that the caveat survives the Better Decision Argument—or will survive for long. Perhaps, some normative quandaries are simply irreconcilable because the existence of multiple values leads to various incompleteness problems.[213] Or perhaps, due process concerns and other constraints simply exclude the possibility.[214]

But subject to those constraints, it would seem that the same Better Decision Argument that applied to simpler technologies applies to these: query whether the machine opinion is better; compare the costs and types of error; consider that due process and similar rights may apply differently in the age of machines; and revise *those* principles as necessary to ensure that the right humans are involved in the normative decision-making before the machine is released "in the wild" and to confirm its performance. In short: decide how we want the machine to decide, and get on with it.

In any event, I hope I have sufficiently sketched the Better Decision Argument at which I take aim. We needn't worry about the caveat just yet,

---

[https://perma.cc/3GJG-N8P6] (discussing open-source legal benchmarking effort). *See generally* Grimmelmann et al., *supra* note 39 (arguing against generative interpretation). There are also a host of challenges to evaluating LLMs for moral performance, competence, and generated "moral reasoning" (i.e., generated text that mimics competent moral reasoning). *See, e.g.*, Julia Haas et al., *A Roadmap for Evaluating Moral Competence in Large Language Models*, 650 NATURE 565 (2026).

211. *See* Liu & Li, *supra* note 87, at 238 (examining the use of AI in Chinese courts); *see also* Ross v. United States, 331 A.3d 220, 229–30 (D.C. 2025) (Howard, J., concurring) (noting majority's and dissent's use of AI in drafting opinions); Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1221–25 (11th Cir. 2024) (Newsom, J., concurring) (discussing Court's use of AI in addressing statutory interpretation).

212. *See, e.g.*, Brożek et al., *supra* note 2, at 435; Ruzica Piskac & Scott J. Shapiro, Leibniz's Dream: How to Automate Legal Reasoning (unpublished manuscript) (on file with author).

213. *See* Cox, *Super-Dicta*, *supra* note 21, at 1646–50; JOSEPH RAZ, THE MORALITY OF FREEDOM 328–35 (1986); Ruth Chang, *Hard Choices*, 3 J. AM. PHIL. ASS'N. 1, 10–11 (2017) (analyzing "what makes a choice hard"); Huq, *Human Decision*, *supra* note 2, at 686 ("[T]he most sophisticated of [the arguments that have entertained the prospect of automated adjudication], an argument for machine-tooled 'micro-directives' forcefully pressed by Anthony Casey and Anthony Niblett, assumes that the law pursues a single goal (such as efficiency) and lacks normative multi-criteriality.").

214. *See* Huq, *Human Decision*, *supra* note 2, at 619 (suggesting a well-calibrated machine would "fold[] in due process, privacy, and equality values"). *See generally* Huq, *Constitutional Rights*, *supra* note 8 (providing due process framework).

and will later return to a stronger way of grounding it in that normative complexity.[215]

B. What Is Better?

It is relatively easy to summarize how normative uncertainty undermines the Better Decision Argument. Quite simply, it is cold comfort to be told you have a right to a "better" decision—and that a "well-calibrated" machine might dutifully provide it—when there is a great deal of uncertainty about what "better" or "well-calibrated" is.

Machines cannot make normative choices, and so machines cannot help you sort that out.[216] Remember: normative uncertainty can be thought of as the kind of uncertainty that your self-driving car could not help you with, the question of when and whether you thought the car *should* turn—under what conditions, by what logic, with what risk.[217] But while you remain unsure, the machinists have settled on an algorithm—they've just picked.[218]

On the other hand, it might seem unclear how normative uncertainty undermines Better Decision Arguments. For one, the problem of normative uncertainty might not seem new; surely this is just the political difficulty of disagreement and the technical problem of value alignment. For another, the solution to it might seem obvious: stop playing philosopher already, think like a programmer, and just pick.[219]

But it turns out that the problem of normative uncertainty is neither the problem of disagreement nor the problem of value alignment. And the solution presumed by the standard framework—just pick—is a nonstarter.[220] I will address the first issue—the nature of the problem—in Subsection 1 and the second issue—the presumed solution—in Subsection 2. Then, in Subsection 3, I will give one example of how recognizing normative uncertainty can revive humanist arguments against the Better Decision Argument. Our discussion will help elucidate why normative uncertainty has gone unrecognized and unnamed, enabling humanists to better articulate

---

215. *See infra* Part III.

216. At least, not if you don't ask them to. *See infra* Section III.B.

217. *See supra* note 8 and accompanying text.

218. *See* Huq, *Human Decision*, *supra* note 2, at 646.

219. Or, at least, pick until the next update or product launch. *See* Cox, *Non-Herculean Data*, *supra* note 94, at 3–4 ("Luckily for the developer, they can revisit their initial selection [in developing and selecting a model], even if at some cost.").

220. Cox, *The Uncertain Judge*, *supra* note 4, at 745.

why machinists' Better Decision Arguments fall flat and do not defeat their claims.

1. What Normative Uncertainty Is Not

The problem of normative uncertainty might not seem new because it is commonly conflated with two related problems: disagreement and value alignment.[221] But these problems are distinct, and so may require different solutions.[222]

The first problem, disagreement, is widely recognized in the legal literature.[223] Many people disagree about the correct resolution to the trolley problem. Although many think it "obvious" that one should turn to save the greater number, others struggle to explain this, and still others believe it is never permissible.[224] There are many who believe the question is poorly formed, and that real trolley problems—which occur under conditions of uncertainty—require a different approach entirely.[225]

But while "[s]ome AI ethicists have seized on moral uncertainty to address moral disagreement in system design," the problems are distinct.[226] Even where there is a sole decisionmaker—even if it were left to each individual user of a self-driving car which mode to select—they might still experience normative uncertainty about which rule to choose.[227] Or, to take our judge, even if she sits alone on a district court, she might be uncertain about how the necessity defense applies in the case of *Primo v. Turner*.

The second problem is the problem of value alignment: "a technical problem about the difficulty of correctly *articulating* the intended rule—of giving the correct instructions or objective."[228] It is a topic of active study

---

221. Cox, *Non-Herculean Data*, *supra* note 94, at 2–3.

222. *Id.* at 3.

223. *See, e.g.*, William Baude & Ryan D. Doerfler, *Arguing with Friends*, 117 MICH. L. REV. 319, 323–28 (2018); Harry T. Edwards, *The Effects of Collegiality on Judicial Decision Making*, 151 U. PA. L. REV. 1639, 1649–51 (2003); Kornhauser & Sager, *supra* note 41, at 11–17 (discussing doctrinal paradoxes on collegial courts).

224. Thomson, *supra* note 5, at 206–10; Foot, *supra* note 5, at 5; Awad et al., *supra* note 49, at 59–64.

225. BARBARA H. FRIED, FACING UP TO SCARCITY: THE LOGIC AND LIMITS OF NONCONSEQUENTIALIST THOUGHT 28–33 (Peter Momtchiloff ed., 2020).

226. Cox, *Non-Herculean Data*, *supra* note 94, at 2 (citing Foot, *supra* note 5; Martinho et al., *supra* note 53).

227. And this is setting aside the normative choices made in terms of system design, both as to *who* gets to make the choice, how, and which options are presented.

228. Cox, *Non-Herculean Data*, *supra* note 94, at 2.

within the AI literature, and—I suspect—the source of some skepticism in the legal literature about the use of machines in certain applications.[229] For example, the Better Decision Argument has also been called the "well-calibrated decision" argument. And what is a well-calibrated decision if not one that has been improved with respect to value alignment? And that's a technical problem, though it is frequently offered as a normative constraint.[230]

Here is an easy way to understand it:

> [C]onsider an example from a popular children's television show: The children tell 'Daddy Robot' to make it so that they never need to clean their playroom again. Daddy Robot reasons that if there were no children, the playroom would never get dirty, and so they would never need to clean the playroom again. He picks up the children and puts *them* in the trash.[231]

What this example shows is the problem of value-alignment, not normative uncertainty. The children "know what they want the machine to do—to keep the playroom clean. But they have fallen prey to the value-alignment problem: they give Daddy Robot instructions [or an objective function] that do not align with what they want done."[232]

These problems are distinct, but "[i]t is not surprising" that they "are often conflated."[233] They are frequently "fellow travelers: If the answers to hard normative questions are not obvious, there is likely to be disagreement. Similarly, if you have normative uncertainty—if you're not sure what the right rules, objectives, or values are—then it will likely be difficult to provide instructions that align with them."[234] Part of my objective in this Article is to expand the language that we have available to diagnose the problems at issue, in the hopes that it enables us to think more critically about how to address them—and to understand what the debate is really about.[235]

---

229. *See* RUSSELL & NORVIG, *supra* note 32, at 33–34 (summarizing the state of technical research); *see also, e.g.*, Huq, *Human Decision*, *supra* note 2, at 687.

230. *See, e.g.*, Huq, *Human Decision*, *supra* note 2, at 652.

231. Cox, *Non-Herculean Data*, *supra* note 94, at 2.

232. *Id.*

233. *Id.* at 3.

234. *Id.*

235. *See infra* Part III.

2. Why the Standard Model Fails

The other reason it might seem unclear how the problem of normative uncertainty stands to recalibrate the humans v. machine debate is that the solution to the problem might seem obvious. Indeed, to the extent other scholars have recognized the problem of normative uncertainty in automated adjudication—though not in so many words—they often frame the issue as about the limits of machines (value-alignment) or the need to make a decision (disagreement), and then proceed to stake out some favored position about how the issue should be resolved, about how to calibrate the machine (and whether such calibration is possible).[236] Their focus is thus on resolving the uncertainty. This strategy follows the "standard model" of AI development, in which rules are either pre-programmed or, for example, a machine-learning model is given an objective function.[237]

In many ways, I think this misplaced focus lies at the core of our unease with machine adjudicators and advisors. We think we must pick. That is, when using machines, we think we must either pick our favored answer, about which we may be uncertain, or else have the machine make the decision, according to our favored approach, or what it gleans about our favored approach from the data and the training. We worry that the machine won't get it right in the individual case, because either we or it will follow a favored approach—even though we think the favored approach might be wrong.

Happily—or unhappily, depending on your perspective!—the standard conclusion is a mistake. We cannot ignore our normative uncertainty in thinking about automating adjudication.

To see why, consider a (hypothetical) machine called "Chief Justice Robot."[238] Chief Justice Robot is an automated decisionmaker that resolves legal disputes on behalf of the state. He takes briefs as inputs and then outputs a "ruling": a decision about whether to grant the requested relief, coupled with an opinion that "explains" how and why the decision was reached.[239] Chief Justice Robot is thus fully autonomous in the colloquial sense of that phrase: although created and maintained by humans, a human

---

236. Kyle Bogosian, *Implementation of Moral Uncertainty in Intelligent Machines*, 27 MINDS & MACHS. 591, 603–05 (2017); Martinho et al., *supra* note 53, at 216–17.

237. RUSSELL & NORVIG, *supra* note 32, at 34 (lamenting that "almost all AI research to date has been carried out within the standard model" but noting "early results within the new framework"). I will return to "uncertain AI" in Section III.B.

238. Volokh, *supra* note 165, at 1182.

239. Cox, *Super-Dicta*, *supra* note 21, at 1581.

is not "in the loop" when a decision is rendered in response to a particular dispute.[240]

I do not know whether such a machine is possible, or even possibly possible. But it is at least imaginable: Some users have employed ChatGPT this way, submitting to it actual briefs, and were convinced (at least, initially).[241] And some legal scholars have argued that that is enough.[242] Meanwhile, China has a smart court system that reportedly does this,[243] and rumors of Estonia employing such a technology are plausible enough to be widely cited (if disputed).[244]

I'll assume for now that Chief Justice Robot is rule-based.[245] This type of design is almost certainly a nonstarter for a variety of reasons,[246] but I use it

---

240. *Cf.* Huq, *Human Decision*, *supra* note 2, at 651 ("If a right to a human decision is to have meaningful content now, therefore, it must be understood to require human judgment at a particular moment: after a machine-learning instrument in the wild encounters and classifies a human actor."). I set aside whether humans might be available to review a machine decision—on analogy to an appeal—because of complications in how the standard of review might impact whether such post-decision review counts as a "human decision." *See id.* at 662 (defining "human decision" as requiring human input "after" deployment); Kaminski & Urban, *supra* note 2, at 2031 (arguing in favor of a "meaningful right to contest AI").

241. *See* Unikowsky, *Trust*, *supra* note 210; Unikowsky, *Trust II*, *supra* note 210 ("The results were otherworldly. Claude is fully capable of acting as a Supreme Court Justice right now."); Grimmelmann et al., *supra* note 39, at 8 ("When proponents like Adam Unikowsky claim that LLMs are good enough for judges to use *now*, they are speaking in the present tense and the indicative mood.").

242. *See* Volokh, *supra* note 165, at 1177.

243. *See* Rachel E. Stern et al., *Automating Fairness? Artificial Intelligence in the Chinese Courts*, 59 COLUM. J. TRANSNAT'L L. 514, 515 (2021).

244. It appears to be fake news. *Compare, e.g.*, Eric Niiler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, WIRED (Mar. 25, 2019), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so, *with* Maria-Elisa Tuulik, *Estonia Does Not Develop AI Judge*, REPUBLIC EST.: MINISTRY JUST. & DIGIT. AFFS. (Feb. 16, 2022), https://www.justdigi.ee/en/news/estonia-does-not-develop-ai-judge [https://perma.cc/KJ75-DGAK] ("As there have been a lot of questions relating the topic of AI Judge, we have to explain that the article about Estonian project of designing a 'Robot/Judge' in Wired from 25th of March 2019, is misleading. There hasn't been that kind of project or even an ambition in Estonian public sector."); *see also* Marco Fabri, *From Court Automation to e-Justice and Beyond in Europe*, 15 INT'L J. CT. ADMIN 7, at 2 (2024) ("Since [the 2019 Wired article] this example has frequently been cited and continues to be hailed as one of the most adavanced [sic] AI projects in the European justice systems. This is simply fake news.").

245. *See* Volokh, *supra* note 165, at 1161.

246. *Cf.* RUSSELL & NORVIG, *supra* note 32, at 24 ("It turned out to be difficult to build and maintain expert systems for complex domains, in part because the reasoning methods used by the systems broke down in the face of uncertainty and in part because the systems could not learn from experience.").

to avoid complexities in exposition that come with data-driven systems and "machine learning."[247]

According to the standard conclusion, we need to decide on the appropriate set of rules governing what the fully automated judge ought to do—we need to settle on what I have elsewhere termed a "jurisprudence."[248] A jurisprudence is simply a theory about what a judge ought (judicially) to do when issuing a ruling. And programming a rule-based machine to follow one will require rendering it complete across the domain of possible cases, even if it is indeterminate.[249] That is, it must always provide an answer of what the judge ought to do, even if that answer is: "pick randomly between these options."[250]

In this way, we need far more information to program Chief Justice Robot than the information that mere mortal human judges have to work with. Most human judges do not have a fully worked out jurisprudence. The domain is too complex. And this difficulty is at least one of the reasons why purely rule-based intelligent agents were doomed to fail (and derisively referred to as "GOFAI"!).[251]

But complete jurisprudences provide a useful modeling tool: one might think of the normative uncertainty a judge faces as uncertainty between two (or more) jurisprudences.

The standard conclusion, restated in terms of jurisprudences, is to just pick the most plausible jurisprudence. Or in machine talk, to just pick the most plausible jurisprudential algorithm.

Unfortunately, the standard conclusion commits two mistakes so long as we have any uncertainty about *which* jurisprudence to select: first, the evidence does not justify proceeding as though your favorite jurisprudence is correct. And second, doing so assumes—implausibly—that all theories say the same thing about the cost of error in all cases. We can see this through two simple examples.

First, the standard conclusion violates dominance. Consider Case I.[252] If, following the standard conclusion, Chief Justice Robot were programmed to

---

247. Yes, I see you thinking that! We'll get there.

248. *See* Cox, *The Uncertain Judge*, *supra* note 4, at 741–44, 760–62; *see also* Cox, *Super-Dicta*, *supra* note 21, at 1586–87.

249. Cox, *Super-Dicta*, *supra* note 21, at 1586–87.

250. *Id.*

251. RUSSELL & NORVIG, *supra* note 32, at 24 (describing the failure of "good old-fashioned AI" or "GOFAI").

252. These cases first appeared in Cox, *The Uncertain Judge*, *supra* note 4, at 782–84. *See also* Cox, *Non-Herculean Data*, *supra* note 94, at 4.

follow Jurisprudence 1, Chief Justice Robot might rule in favor of the defendant (Δ)—even though there is a significant chance that doing so would be a mistake. If Chief Justice Robot instead took the normative uncertainty into account, Chief Justice Robot would secure his objective—of doing what is judicially right—by ruling in favor of the plaintiff (π).

Figure 4. Case I

	Jurisprudence 1 ($p \geq .51$)	Jurisprudence 2 ($p \leq .49$)
π	✓	✓
Δ	✓	✗

Second, the standard conclusion fails to take into account the stakes. Consider Case II. Different jurisprudences often disagree not only about how to decide a case, but also about how much it matters to get it right. If Chief Justice Robot were programmed to follow only Jurisprudence 1, and to disregard the recommendations of Jurisprudence 2, then in deciding Case II, Chief Justice Robot will rule in favor of the defendant. But doing so involves a significant risk of making a huge mistake for the benefit of avoiding only a minor one.

Figure 5. Case II

	Jurisprudence 1 ($p \geq .90$)	Jurisprudence 2 ($p \leq .10$)
π	✗ Minor mistake	✓
Δ	✓	✗ HUGE MISTAKE

Quite simply, the standard conclusion is a huge mistake. We are right to feel a certain amount of unease with automated adjudicators, especially

those based on the standard framework. It is not merely disagreement or the value-alignment problem. It's that if we're uncertain about those values, we will be fools to create a Chief Justice Robot who's not.[253] It is a mistake to attempt to hardwire Hercules, when we do not know what Hercules should do.[254]

3. Reviving the Legal Subject

Having shown a fundamental problem with the machinists' Better Decision Argument, it may be helpful to illustrate, more concretely, how normative uncertainty provides the humanists with another tool for defending humanist principles against the Better Decision Argument. Here I will consider just one: a group of humanist arguments that often appear under the heading of "Legal Subjects" or calls for an "Individualized Decision." Call these the humanists' "Individualized Decision Arguments." Professor Huq suggested that "a subtler approach is necessary to make sense of [them]."[255] Normative uncertainty can provide that.

---

253. At this point, a computer scientist might think some version of: OK, so what's the meta-algorithm? It turns out that *solving* normative uncertainty is not trivial. In brief, the difficulty is that different theories about what one ought to do might not be comparable. For example, to apply a common approach to empirical uncertainty, namely expected value theory, you would need the jurisprudences to both admit of degrees (i.e., provide a cardinal ranking) and for those degrees to be on the same scale (i.e., co-cardinality). But in the case of normative uncertainty, that is often exactly what is in dispute: you can't count on the jurisprudences agreeing on a scale or even having a scale at all. For discussion, see Cox, *The Uncertain Judge*, *supra* note 4, at 797–802 (explaining difficulties related to intertheoretical comparison); *see also* Cox, *Super-Dicta*, *supra* note 21, at 1579 n.7 (noting that one of the leading solutions to the problem in the moral case was recently called into doubt) (citing Johan E. Gustafsson, *Second Thoughts About My Favourite Theory*, 103 PAC. PHIL. Q. 448, 451–52 (2022)); MACASKILL ET AL., *supra* note 4, at 214–15 (noting that leading philosophers believed the problem might be intractable).

254. Dworkin, *Hard Cases*, *supra* note 138, at 1108–09.

255. Huq, *Human Decision*, *supra* note 2, at 678 ("Although I am not convinced that it yields a general objection to machine decisions, I think that with certain assumptions and under certain conditions, it can be deployed to resist specific substitutions of machine for human decisions."). Professor Huq narrows the Individualized Decision Argument to cases where "the human decision maker will have access to individualized evidence, whereas the machine decision maker would have access only to statistical, or population-wide, information." *Id.* He then argues that it might be used to resist machine decisions in cases where deterrence matters but expresses doubt about the narrowed argument's reach and strength. *Id.* at 679–80. With normative uncertainty, humanists need not accept such narrowing or "probably flawed" assumptions about a machine's access to individualized information. *Id.* at 680.

The rough shape of these arguments is as follows: machines should not replace human decisionmakers in contexts like adjudication because machines treat legal subjects not as individuals, but as statistics.[256] This treatment is a problem for many reasons, or so humanist Individualized Decision Arguments suggest.[257] For one, such treatment fails to take an individual's characteristics seriously.[258] For another, such treatment deprives the individual of an opportunity to challenge the algorithm, both facially and as applied—aptly illustrated by a veritable parade of machine-decisions-gone-wrong, from Catherine Taylor who was denied federal housing assistance after an "automated, 'webcrawling[,] data-gathering' background check" erroneously flagged her file, to Tammy Dobbs who lost full medical assistance until a lawsuit uncovered her benefit was erroneously cut in half because the algorithm ignored one of her many health conditions.[259] In addition to known horribles, humanists argue, machines also miss important features because they are bad at extrapolation: if a feature is not in the training data or otherwise encoded for, the machine will miss it, no matter how important to an individual's case it might be.[260] And finally, machine decisions deprive legal subjects from appealing to a human for their humanity: for a discretionary or

---

256. *See id.* at 676 ("Roughly stated, the intuition here is that the state should take action against a person solely on the basis of their own behavior or merits. It should treat them, that is, 'as an individual.'").

257. *Id.* at 675–76.

258. *Id.* at 676; *see also* Katrina Geddes, *The Death of the Legal Subject*, 25 VAND. J. ENT. & TECH. L. 1, 50 (2023) (observing, without "seek[ing] to answer the question of whether predictive algorithms should or should not be used in judicial decision making," that "the elevation of algorithmic knowledge represents a transfer of narrative power from the individual (whose behavior is being predicted) to the data capitalist (whose prediction now forms the basis for decision making)").

259. *See, e.g.*, Huq, *Human Decision*, *supra* note 2, at 616 (quoting CATHY O'NEIL, WEAPONS OF MATH DESTRUCTION: HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY 152–53 (2016)) (explaining that the error in Taylor's case was only corrected once "'one conscientious human being' [took] the trouble to look"); Kesari et al., *supra* note 17, at 1 (citing Colin Lecher, *What Happens When an Algorithm Cuts Your Health Care*, VERGE: SCI. (Mar. 21, 2018), https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy); *see also* Citron, *supra* note 1, at 1252, 1256–57, 1276 (listing parade of horribles, from welfare decisions to child support to voter-roll purges to destroyed credit from default judgments).

260. *See* O'NEIL, *supra* note 259, at 25 ("To create a model . . . we make choices about what's important enough to include, simplifying the world into a toy version that can be easily understood and from which we can infer important facts and actions . . . A model's blind spots reflect the judgments and priorities of its creators.").

equitable or merciful exception.[261] In sum, machine decisions prevent individual legal subjects from supplying information that might reveal a mistake, from identifying additional features that might affect the outcome, and from appealing to discretion. Or so some humanist Individualized Decision Arguments go.[262]

The machinist Better Decision Argument had responded to these humanist Individualized Decision Arguments by arguing, in turn: that individualized decisions are often not required,[263] that a well-calibrated machine will correct the errors and so the as-applied arguments "conflate the short-term gain from human review" with "dynamic optimality" in the long term,[264] that mercy and discretion are poor hooks on which to hang a *right* to a human.[265] Indeed, machinist Better Decision Arguments often suggest that the importance of consistency and fairness may counsel *in favor* of machine decisions given how machines can serve as a "pre-commitment device."[266]

Whatever the merits of these various arguments, normative uncertainty has something new to offer the humanists in grounding Individual Decision Arguments. Humanists can now argue that the machinists' Better Decision Argument improperly resolves normative uncertainty about individual cases.

The humanist response begins by acknowledging that machines—if they exhibit the consistency necessary to (seemingly) defeat the Individual

---

261. *See generally* Re & Solow-Niederman, *supra* note 57 (arguing that AI will not only deprive individual legal subjects from appealing to a human for their humanity, but "[b]y offering efficiency and at least an appearance of impartiality, AI adjudication will foster a turn toward 'codified justice,' that is, a paradigm of adjudication that favors standardization above discretion"); *cf.* Geddes, *supra* note 258, at 43–45, 50 (recounting significance of shifting "narrative power" from individuals to "data capitalists").

262. There are many other humanist bases, but my aim here is to illustrate how normative uncertainty can strengthen some humanist arguments in the face of Better Decision Arguments, so we'll focus on these. *See* Huq, *Human Decision*, *supra* note 2, at 662 ("The most plausible gloss of a claim to additional human input needs to hinge on an ex post human role after a machine-learning decision has been delivered. That would mean an individual subject to a machine decision could respond directly to that decision by drawing it to a human's attention. Such a right might respond to the possibility, for instance, that the 'feature values' used to train an algorithm excluded some parameter of relevance to a subset of individuals, but not the general population.").

263. *Id.* at 677 (noting many human decisions also "fail[] to individuate").

264. *Id.* at 663–64.

265. *Id.* at 660–61.

266. *Id.* at 675; Re & Solow-Niederman, *supra* note 57, at 257–58 (noting that availability of machine adjudication may also shift preferences in favor of such "codification").

Decision Argument—will encode "relevant" features.[267] In other words, the machine encodes a jurisprudence and unlike a human judge, consistently applies it.

Next, the humanist holds their fire on the individual and the parade of machine error. Instead, the humanist points to normative uncertainty: there is uncertainty about those encoded features and how they apply—about the correct jurisprudence. Even so, the machinist "just picked."

Now the humanist goes in for the kill: by "just picking," machinists commit rational error. By claiming that you can "just pick" your way to a "well-calibrated," consistent machine, machinists mistakenly assume that different jurisprudences agree about the stakes in all cases. But sometimes jurisprudences disagree about the stakes, and it matters.[268] If two legal subjects differ along a feature deemed relevant by a non-encoded but plausible jurisprudence, the rational calculus might very well tip about how the individual's case should turn out.[269] That is, it is a case where a human decisionmaker should rationally follow a *different* jurisprudence instead. But a machine cannot do so because a favored jurisprudence is already encoded.[270] Similarly, if two legal subjects differ along a feature deemed relevant, but about which there are differing views on how the feature is applied, that might change the rational calculus. Or perhaps, a legal subject reveals a new feature that should change the decisionmaker's credences about which jurisprudences are plausible—about which approach the machine should have been encoded with in the first place.[271]

---

267. Huq, *Human Decision*, *supra* note 2, at 675. Such encoding is also a problem for machine-learning systems under the traditional model. *See* Lehr & Ohm, *supra* note 2, at 657 ("[W]ith a running model, all we can do is rue the choice that has already been made."); *infra* Section III.B.

268. *See, e.g.*, Cox, *The Uncertain Judge*, *supra* note 4, at 791–96 (providing a decision-theoretic reconstruction of *Brown v. Board of Education*, 347 U.S. 483 (1955)); *see also* Cox, *Super-Dicta*, *supra* note 21, at 1601–08.

269. *See* Cox, *The Uncertain Judge*, *supra* note 4, at 781–85.

270. *See id.* at 776–86.

271. Recall that a decisionmaker's credences are the strengths of her beliefs in different propositions, like that a case is correctly decided or that a given jurisprudence is the correct account of what a judge ought (judicially) to do. *See* Cox, *Super-Dicta*, *supra* note 21, at 1605 ("[A] judge's epistemic credences about jurisprudences will be consistent with her credences about how particular cases should be resolved, about how they can be made consistent with past cases, about whether past cases were wrongly decided, about what should be done about that (if anything), and so forth. A judge's actual credences almost certainly diverge from her epistemic credences—a large part of reflecting and thinking and experience is attempting to render them consistent!").

Machines do not currently take normative uncertainty into account.[272] But perhaps we have an instinct that wise judges do. This is what it is to take the individual seriously.[273] To get a machine that acts like Hercules, you cannot attempt to encode him directly when you don't know what he should do. A "well-calibrated" machine decision is often an uncertain one, and as we shall see, that brings humans back into the loop.[274]

III. RECALIBRATING THE DEBATE

Parts I and II showed how normative uncertainty seems to undermine the leading argumentative strategies on both sides of the "right to a human decision" debate. Because of normative uncertainty, humanists' Arguments from Explanation do not count in favor of humans in precisely those high stakes contexts, like adjudication, where explanations for decisions are owed.[275] Meanwhile, machinists' Better Decision Arguments commit critical errors in both ignoring the problem of normative uncertainty and offering (or assuming) a nonstarter of a solution.[276] Normative uncertainty thus seems to have deflated the strongest argument in each "camp."

But having done that, where do we go from here? How do we recalibrate the debate? In this Part, I will offer two suggestions.

First, I aim to sharpen the question at the core of the human/machine debate, about the timing and manner of human involvement in decisions in particular cases.

Second, I offer some parting thoughts on a novel use for machines in adjudication: to assist judges in grappling with normative uncertainty, a practice that early indicators suggest may already be underway.[277] Contrary to what many in the debates assume, one of the best human-machine partnerships may be on explicitly normative questions. And it is a use where

---

272. *But see infra* Section III.B. Do some predictive machines account for normative uncertainty by aiming to predict what humans would do—humans that take normative uncertainty into account? Possibly, but it is not clear such machines would do so in a principled way, and it appears that the risk of distortionary effects is significant. For discussion, see Cox, *Non-Herculean Data*, *supra* note 94, at 2, 7–9. *See also* Cox, *Super-Dicta*, *supra* note 21, at 1639–41.

273. This argument resonates with Ruth Chang's arguments about how parity creates space for agency by enabling normative *choice*. *See* Chang, *supra* note 213, at 17–18; *see also* Shiffrin, *supra* note 45, at 1217; Stone, *supra* note 45, at 118; Bagchi, *supra* note 45, at 541.

274. *See infra* Section III.B.

275. *See supra* Section I.B.

276. *See supra* Section II.B.

277. *See infra* notes 327–328 and accompanying text.

concerns about explanation and transparency of *individual* decisions matter less, such that global explanations of the sort provided by xAI will do.

A. A Better Question

Having undermined the leading arguments in the "right to a human decision" debate, the question arises: Where does that leave us? Some recalibration is clearly in order.

In what follows, I will suggest that the debate be reoriented from focusing on whether a machine produces "better" decisions to focusing on the appropriate scope of choice for decision-making under conditions of normative uncertainty—for deciding when we have uncertainty about what "better" is.[278] "Scope of choice" is, roughly, how much to decide at once—like ordering from a set menu, following a diet, or ordering a la carte.

To get there, I'm going to first relax some of the simplifications I have employed thus far. Then, I will turn to thinking about solutions to normative uncertainty. To be clear, I still do not have *the* solution to the problem of normative uncertainty.[279] No one does.[280] But the only way we will get from here to there is by sharpening the questions that we need to ask.

Let's begin by relaxing the simplifications. Throughout, I have taken some obvious liberties in describing a "two-sided" all-or-nothing debate about a "right" to a human decision. I did so because I wanted to evaluate argument types: I wanted to explore what certain types of arguments could or couldn't show about whether there are *categorical limits* on replacing humans with machines in certain contexts. That terrain was a little

---

278. Whatever the flavor of normative uncertainty, including epistemic—i.e., about what justifies belief. *Cf.* Fagan & Levmore, *supra* note 25, at 30 (discussing difficulty of knowing when to overrule AI due to reversal paradoxes).

And yes, there may be normative uncertainty about the rational ought! And that may cause concerns about regress. Chidi from The Good Place—that send-up of philosophers—was incorrigible for a reason. That said, there is reason to think that regress is not the problem supposed. Different meta-algorithms may converge. But that is work for another day. *See* MACASKILL ET AL., *supra* note 4, at 30–33.

279. Sorry.

280. So, sorry not sorry. On the difficulty of the problem, see, for example, Gustafsson, *supra* note 253, at 448, 451–52 (undermining view author previously advocated); Cox, *Super-Dicta*, *supra* note 21, at 1579 n.7 ("To wit, an argument in favor of one of the leading solutions for the analogous (and arguably easier) moral problem was recently undermined by one of its (former) proponents." (citing Gustafsson, *supra* note 253, at 451–52)); Cox, *The Uncertain Judge*, *supra* note 4, at 797–802 (explaining difficulties with intertheoretical comparison).

complicated, so it helped to simplify where we could. But the debate is not so neat.

For starters, (most) everyone realizes that machines are not yet *wholly* free of human judgement. Rather, choices about how to design and deploy machines will affect the shape and content of downstream "machine" decisions.[281] "It's humans all the way down."[282] And so, the more nuanced way to understand the debate—so most say—is as one about the timing and manner of human involvement in decisions in particular cases.[283]

For another, the debate does not divide neatly into two camps, with each insisting that machines or humans win in all contexts. Rather, individual scholars may "switch" sides depending on the context, pro-machine in some, pro-human in others.

So now, we can be more nuanced and ask: when should we switch sides? That is, if the core question is not really about a "right to a human decision," but about the timing and manner of human involvement in decisions in particular cases, then what should inform that choice? Or, to flip the script, when is a machine an appropriate tool?

We have seen that the machinists' Better Decision Arguments have framed this question in terms of calibration: Can the machine be tweaked to better instantiate a certain set of values than is available from a human decision? But as we have seen, Better Decision Arguments were misguided. Calibration often forces trade-offs, and we don't know which to make.[284] Mitigating one type of error may give rise to other types of error: sometimes fast response times are necessary to avoid a crash; other times, you jump the gun and cause one. Sometimes reducing false negatives means increasing false positives, and vice-versa. How can you calibrate a machine when calibration requires making difficult trade-offs about which you are uncertain?[285] The machinists' Better Decision Arguments tend to confuse

---

281. Crootof et al., *supra* note 25, at 443 (collecting literature) ("It's humans all the way down. . . . We are not the first to note this.").

282. *Id.*

283. Huq, *Human Decision*, *supra* note 2, at 651; Kaminski & Urban, *supra* note 2, at 1972.

284. *See, e.g.*, Crootof et al., *supra* note 25, at 477; *see also* Cox, *The Uncertain Judge*, *supra* note 4, at 761–62 (arguing that normative uncertainty arises in part because jurisprudences are not infinitely malleable and so trade-offs must be made); Cox, *Non-Herculean Data*, *supra* note 94, at 2–3 (comparing the difficulty the judge faces with that of programmers).

285. *See supra* Section II.B.1.

this difficulty with other problems, and the implication—just pick—commits not one, but two, rational errors.[286]

But while the machinists were wrong that you should "just pick," they were not wrong that you still need to decide despite your uncertainty.

So, the next question is: how much should you decide? That is, what is the "scope of choice"? Are you dictating a weekly meal plan, leaving it up to the babysitter, or ordering daily à la carte? Are you setting a rule or process that will dictate outcomes, or leaving room for some flexibility in the moment?

Machines, at least the "standard model," force an answer to this question of scope: the machine's scope of choice is broad.[287] And this is true on the standard AI model regardless of whether you employ good old-fashioned rule-based AI, or a newer data driven model with an objective function. The mechanism may differ: either you choose rules that dictate how individual encounters would be resolved (a meal plan with some contingencies baked in—"order pizza if running late"),[288] or you pick a model, give some instructions about your goal (an "objective function"), and delegate, occasionally fine-tuning through feedback and training (as when you pick a babysitter and shout "healthy but tasty" as you escape for the weekend).[289] But either way, you have elected to resolve many smaller, individual decisions through this process—even if it leaves such individual decisions to chance. That is what you do when you elect to use a machine.

Humans do not force this selection of a broad scope of choice. Indeed, humans are notoriously bad at sticking to one.[290] And so, if we understand the debate as not "when human" but as "when machine," we can now more sensibly ask at least one question about when to "switch sides" from humanist to machinist: Is this a context where a broad scope of choice is appropriate for making decisions under normative uncertainty? Is this a context in which flexibility about scope of choice is desirable?

---

286. *See supra* Section II.B.2.

287. RUSSELL & NORVIG, *supra* note 32, at 33–34. More on new models in Section III.B.

288. *Id.* at 23 ("[Expert systems] derived [their expertise] from large numbers of special-purpose rules."); *see also id.* at 24 ("It turned out to be difficult to build and maintain expert systems for complex domains, in part because the reasoning methods used by the systems broke down in the face of uncertainty and in part because the systems could not learn from experience.").

289. *Id.* at 22–27.

290. *See* Kornhauser & Sager, *supra* note 41, at 53–56.

Different scopes of choice will be appropriate in different contexts. And selecting a scope of choice is a terribly difficult problem.[291] Thinking through how to do so is thus work for another day. But I would like to illustrate the kinds of considerations that inform the selection of scope of choice, both to suggest a path forward and because they provide a new lens for understanding the debate.

First, some in the literature seem to assume that the selection of machine versus human turns on the complexity of the normative judgments to be made.[292] On the one hand, normative uncertainty would seem to explain this: the more complicated the terrain, the more frequently individual encounters will present "hard" cases, the greater the normative uncertainty. And often, when selecting a scope of choice under conditions of uncertainty, it is rational to choose narrowly—to wait to fill in the details.[293]

But this is not always the case. Sometimes scope of choice is informed instead by what you're trying to achieve. Return to driving. Machines may force a broad scope of choice. But human drivers need to employ a broad scope of choice too—remember the Squirrel Rule.[294]

Second, sometimes it is better to avoid real-life trolley problems in the first place than to solve your uncertainty about what to do once you're in one. Some scholars have critiqued the obsession with the trolley problem in self-driving cars on this basis.[295] Avoiding trolley problems may be more difficult than appreciated.[296] But the point stands: one reason to prefer a broad scope of choice is if machines can reduce the trolley problems that generate uncertainty in the first place.

---

291. *Cf. id.* at 57–59 (summarizing arguments that the appropriate decision protocol for collegial adjudication varies).

292. Huq, *Human Decision*, *supra* note 2, at 636 ("Machines cannot (yet?) resolve the difficult and inescapably normative questions of aggregation, distribution, and belonging that characterize politics.").

293. *See* JOHN RAWLS, A THEORY OF JUSTICE 360 (1999) ("A plan will, to be sure, make some provision for even the most distant future and for our death, but it becomes relatively less specific for later periods. . . . Indeed, one principle of rational choice is postponement."); *cf.* SHAPIRO, *supra* note 108 (introducing planning theory of law).

294. *See supra* Section II.A.

295. Kamm, *supr*a note 7, at 88; *see supra* Section II.A.

296. *See* Final Judgment at 1–2, Benavides v. Tesla, Inc., No. 1:21-cv-21940, (S.D. Fla. Aug. 4, 2025) (entering judgment on $242,570,000 verdict against Tesla for design defect causing crash with human-in-the-loop crash avoidance system). *See generally* Crootof et al., *supra* note 25, at 438 (describing underappreciated difficulties of integrating machines into human processes); Charlotte A. Tschider, *Humans Outside the Loop*, 26 YALE J. L. & TECH. 324, 383 (2024).

Conversely, sometimes you need to confront the trolley problem: hard cases are not bad cases even if they supposedly make bad law. And so, it would be a dereliction of duty to address normative uncertainty by trying to avoid it. For example, in law, judges should not necessarily discourage parties from bringing hard cases.[297] In such situations, a narrow scope of choice may be preferable because a narrower scope of choice is usually less informationally demanding, and there may be less normative uncertainty about a particular case than a set of them.[298]

Third, new encounters bring new information, the likelihood and significance of which may vary by context. This problem is frequently noted in the literature.[299] But it is usually addressed in the context of calibration and accuracy: the machine does not make the right call in a new encounter because the new encounter exhibits features that were missing either from the rule or the data set.[300] That is, the new information reveals that the machine was not value-aligned—it does not perform as intended.

But there is a further reason that new information can be significant: it can reveal unappreciated normative uncertainty. This is the game that is played by trolleyology: As soon as you articulate a principle for your intuition in one case, new features are added or the features are changed.[301] But your principle yields implausible results in this new case. And so now, you need to figure out a principle that reconciles them. Sometimes you can't! Aha, uncertainty.

This "bug" of trolleyology is why it presents such a good example for discussing normative uncertainty. If the new features—the added context—of the second case change our assessment of the principle we articulated in the first case, then this suggests that there's normative uncertainty about those principles.[302]

---

297. *See* Nelson & Schwartzman, *supra* note 151, at 1112–14 (critiquing constitutional least-cost avoider theory); Cox, *Super-Dicta*, *supra* note 21, at 1650 (noting that copyright doctrine incentivizes hard cases, with reason).

298. One of the reasons that normative uncertainty is so acute in the case of judging is that the reason-giving practice forces a broader scope of choice than merely issuing a decision. *Cf.* Schauer, *supra* note 82, at 658–59 (discussing the tension between reason-giving and case-by-case determinations, since giving reasons typically commits a decisionmaker to principles more general than the particular decision).

299. *See, e.g.*, Kaminski & Urban, *supra* note 2, at 1990.

300. *Id.*; Deeks, *supra* note 11, at 1835; Berman, *supra* note 72, at 1288–90.

301. Now consider organs!

302. If you thought the point of trolleyology was to actually find rules and principles, you might find this context bait-and-switch infuriating. Indeed, it is a much-maligned feature of trolleyology. But that is to miss the real lesson of trolleyology, which is that ethics is hard.

And so, one relevant consideration is: what are the stakes likely to be in that future case? Is there convergence in the normative principles about how bad it would be? Are those offset, in any way, by the other cases which could be gotten "right" by the use of a broader scope of choice? In a nutshell, how important is it to brake for kangaroo, and how does that trade-off against a machine's ability to make the right normative call in all the cases with squirrels?[303]

There is much more to be said, but it is work for another day. My point is simply that the better question is not which is better, "human or machine," but what scope of choice would be rational in this context, given normative uncertainty? And how flexible should the selection of scope of choice be?

B. Can Machines Help with Normative Judgment After All?

Before we go, I want to make a further suggestion: that we learn from the engineers about how to think about solutions to these normative difficulties. The funny thing about engineering self-driving cars is that the hardest technical questions relate to the normative ones. Engineers know this. For example,

> [I]n designing a self-driving car, one might think that the objective is to reach the destination safely. But driving along any road incurs a risk of injury . . . ; thus, a strict goal of safety requires staying in the garage. There is a tradeoff between making progress towards the destination and incurring a risk of injury. How should this tradeoff be made? [304]

Computer science conventionally understands this problem to be the "value alignment problem."[305] According to it, "the values or objectives put into the machine must be aligned with those of the human."[306] But unfortunately, "[a]s we move into the real world, . . . it becomes more and more difficult to specify the objective completely and correctly."[307] And the stakes become increasingly high as the technology grows more powerful.[308]

---

303. *Supra* Section II.A; *see also* Kaminski & Urban, *supra* note 2, at 2008.

304. RUSSELL & NORVIG, *supra* note 32, at 5.

305. *Id.*

306. *Id.*

307. *Id.* at 4–5.

308. *See id.* at 33. *See generally* Norbert Wiener, *Some Moral and Technical Consequences of Automation*, 131 SCI. 1355, 1358 (1960) (describing the potential consequences of using machines whose actions may be beyond human control).

Professors Peter Norvig and Stuart Russell colorfully describe this problem as the "King Midas problem," after that "legendary King in Greek mythology" who wished that everything he touched be turned to gold.[309] And Professor Norbert Wiener invoked similarly cautionary tales about genies and wishes.[310]

I think it is telling that this problem is associated with wishes. I said earlier that value alignment and normative uncertainty were fellow travelers. That is true at a deep level.

One moral of these stories, of being "careful what you wish for," is that it is very difficult to correctly articulate your wishes. That is, the moral is about the difficulty of value alignment.

But genie stories have another moral. You might not really know what you should wish for.[311] You should realize that you might be wrong. That is, the moral is also about the problem of normative uncertainty.

Computer scientists have been working for some time on the King Midas problem: on the problem of value alignment, in the shadow of normative uncertainty. And there is much to learn from their efforts.

They have already figured out that "just pick" is a non-starter as a solution.[312] As Professors Norvig and Russell have written, the "standard model" of AI research aims to construct "agents that *do the right thing*," where "the right thing is defined by the objective that [engineers] provide to the agent."[313] But the standard model "is probably not the right model in the long run" because it "assumes that we will supply a fully specified objective to the machine"—an unrealistic and dangerous strategy, as we learned from genies.[314]

The solution? Uncertain AI: "machines that strive to achieve human objectives but know that they don't know for certain exactly what those objectives are."[315]

That is, the engineers have figured out that if you want to build a Herculean machine—that behaves in the way you want—you can't hardwire him directly.

---

309. RUSSELL & NORVIG, *supra* note 32, at 33.

310. Wiener, *supra* note 308, at 1358.

311. There is a third: there are some wishes a genie cannot grant, like love.

312. *Supra* Section II.B.2.

313. RUSSELL & NORVIG, *supra* note 32, at 4 ("This general paradigm is so pervasive that we might call it the standard model.").

314. *Id.*

315. *Id.* at 33.

This new framework is sufficiently new that "almost all AI research" as of 2021 had been done under the standard model.[316] But we may take inspiration from some of the early tools. One of the early results come from assistance games.[317] In such games, a "human has an objective and a machine tries to achieve it, but is initially uncertain about what [that objective] is."[318] "Solutions to assistance games include acting cautiously. . . and asking questions."[319]

Engineers aren't quite right about the link between human objectives and preferences. These games typically take revealed preferences as indicative of human objectives. But there is an increasing awareness that human preferences are not consistent and so "it may not be clear what AI systems *should* be doing."[320] And this increasing awareness is to be encouraged: my revealed preferences often don't align with my consciously chosen objectives, let alone with what my objectives should be.[321]

Naming the problem—"normative uncertainty"—may assist engineers in understanding that the difficulty is not merely one of inconsistent preferences. Rather, there are deep theoretical difficulties within normative theory about how multiple values can generate hard cases, not because preferences are irrational, but because trade-offs are unknown.[322] These theoretical difficulties in turn generate a practical difficulty about how to decide in such cases, a question about which there may be normative uncertainty.[323]

While they don't have the full picture, engineers are—interestingly—focused on the right question: scope of choice. The new technical framework tries to proceed cautiously and to ask questions—a bit like the judicial enterprise. The new approach operates with a narrower scope of

---

316. *Id.* at 34 ("It is perhaps unfortunate that almost all AI research to date has been carried out within the standard model, which means that almost all of the technical material in this edition reflects that intellectual framework.").

317. *Id.*

318. *Id.*

319. *Id.* at 1004.

320. *Id.* at 34.

321. See, for example, my weekly screentime reports from Apple.

322. *See* sources cited *supra* note 213; *see also* Cox, *Super-Dicta*, *supra* note 21, at 1641–49. For these reasons, it is not mere coincidence that the "most sophisticated" argument in favor of machine adjudication "assume[d] that the law pursues a single goal . . . and lacks normative multi-criteriality." Huq, *Human Decision*, *supra* note 2, at 686 (citing Anthony J. Casey & Anthony Niblett, *A Framework for the New Personalization of Law*, 86 U. CHI. L. REV. 333 (2019)) (noting that Casey and Niblett appear to assume that goal is efficiency).

323. Cox, *Super-Dicta*, *supra* note 21, at 1641–49.

choice instead of settling everything in advance. So, contrary to what I wrote above, modern machines might be an appropriate tool even in contexts like adjudication to the extent they are able to operate with a narrow scope of choice. But note that they only do so when they are put *in conversation* with humans.

This conversational assistance game is notable. It has been commonly assumed that machines can't help with moral decisions, except by offering factual predicates or factual predictions—the descriptive, as opposed to normative, inputs to moral decision-making.[324] One reason for this assumption was technical limitations.[325] But another seems to have been the difficulty of the problem.[326]

And yet, now that we have technology that at least *appears* capable of replicating moral discourse with the rough competence of a philosophy undergraduate,[327] judges have turned to it *for assistance in making normative judgments*. To date, they have mostly done so to test empirical or interpretative questions. For example, Judge Deahl from the District of Columbia Court of Appeals recently used ChatGPT to test his intuitions about common knowledge.[328]

This judicial use is striking. And I want to suggest that we learn from engineers that more may be possible. LLMs hallucinate in a way that might not be fixable for source-based legal reasoning or interpretation like Judge Deahl's use.[329] But LLMs can serve as reasonably intelligent

---

324. To my knowledge, there is one exception. *See* Richard Re, *Artificial Authorship and Judicial Opinions*, 92 GEO. WASH. L. REV. 1558, 1572–74 (2024) ("AI can and often will improve judicial deliberation [by, *inter alia*,] brainstorm[ing] arguments and counterarguments.").

325. *E.g.*, Huq, *Human Decision*, *supra* note 2, at 636 ("[M]any decisions related to the law fall outside the domain of the technologically plausible.").

326. *See id.*

327. *See, e.g.*, Leiter, *supra* note 210 (sharing example essay produced in response to prompt shared by philosophy professor Jason Bridges).

328. Ross v. United States, 331 A.3d 220, 236–37 (D.C. 2025) (Deahl, J., dissenting); *see also id.* at 231 (Howard, J., concurring) (commending Judge Deahl and the majority's use of ChatGPT, noting "this was no delegation of decision-making, but instead the use of a tool to aid the judicial mind in carefully considering the problems of the case more deeply."); Fagan & Levmore, *supra* note 25, at 33–34 (proposing using machines to test law's empirical basis of legal rules).

329. Even attempts to use "retrieval-augmented general AI" to solve this problem have had middling results, at best. *See generally* Varun Magesh et al., Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools (May 30, 2024) (unpublished manuscript) (on file with arXiv) , https://arxiv.org/pdf/2405.20362v1 [https://perma.cc/8RT3-5LYM] (providing typology of "hallucinations" in legal context, finding that "while [retrieval-augmented generative AI] appears to improve the performance of language models in answering legal

"interlocutors," noting—erm, predicting—flaws in reasoning, principles that will come into conflict, and unexpected connections between features.[330] Hybrid systems that combine natural language processing with defeasible logic may be even more impressive on this score.[331]

I don't want to end on too Pollyanna-ish of a note. But I want to suggest that there is a need for working through these ideas, that this is one of the ways we address normative uncertainty. I doubt any judge has a decision matrix of the sort described above in mind when making decisions. That is, I doubt any judge thinks quite so mathematically in terms of uncertainty between complete jurisprudences, fully specified across the full range of cases—even if this is a useful way to model less formal thinking.[332] Complete jurisprudences are too informationally demanding for humans. But a machine might be able to help map the construct of our views and organize normative information in a more comprehensive way. This mapping might, in turn, allow for the use of more sophisticated methods for resolving normative uncertainty.[333] It may even change which cases we *perceive* as easy or hard.

There are concerns, to be sure—and such concerns are both serious and numerous. Some concerns are unique to the legal setting, like judicial ethics and what might seem like ex parte communications with a privately designed and managed chatbot.[334] Some concerns relate to how machines

---

queries, the hallucination problem persists at significant levels," and characterizing reasons for some of these failures); Guha et al., *supra* note 210 (presenting a legal benchmarking effort intended to assess and develop large language models for law). For criticisms of generative interpretation, see, for example, Grimmelmann et al., *supra* note 39.

330. For a similar suggestion with respect to use in opinion writing, see Re, *supra* note 324, at 1572–74. For a similar point, but with respect to using pattern recognition to evaluate the empirical support for legal rules, see Fagan & Levmore, *supra* note 25, at 33–34.

331. *Cf.* Piskac & Shapiro, *supra* note 212.

332. Cox, *The Uncertain Judge*, *supra* note 4, at 759.

333. For example, "sophisticated jurisprudences might usefully be modeled using the different computational models" already in existence. Cox, *Non-Herculean Data*, *supra* note 94, at 3 (citing Latifa Al-Abdulkarim et al., *A Methodology for Designing Systems to Reason with Legal Cases Using Abstract Dialectical Frameworks*, 24 A.I. & L. 1, 3–4 (2016); Katie Atkinson & Trevor Bench-Capon, *Practical Reasoning as Presumptive Argumentation Using Action Based Alternating Transition Systems*, 171 A.I. 855, 861 (2007); Katie Atkinson & Trevor Bench-Capon, *Taking Account of the Actions of Others in Value-Based Reasoning*, 254 A.I. 1, 4–5 (2018); and Trevor Bench-Capon, *Value-Based Reasoning and Norms*, *in* 285 22ND EUROPEAN CONFERENCE ON ARTIFICIAL INTELLIGENCE 1664–65 (2016)); *see also* Piskac & Shapiro, *supra* note 212, at 2–3, 7 (using LLMs to translate input into structured data for use with decision-tree formalization of technical areas of law).

334. *See* Ross v. United States, 331 A.3d 220, 230–31 (D.C. 2025) (Howard, J., concurring) (describing scenarios where "confidential judicial deliberative information has potentially

may affect normative reasoning itself: machine strengths (and weaknesses) may make certain modes of reasoning or considerations more or less salient.[335] And there are all the standard concerns for any machine system, like bias and cybersecurity.[336] Other concerns relate to the difficulties inherent with introducing machines into human loops: problems like skill fade, where reliance on tech erodes human skills and renders them less well-suited to handle exception cases.[337] One might be particularly concerned with priming dangers and sycophancy, not only because current chatbots cause some people to become delusional, with devastating consequences, but also for the more mundane reason that sycophancy can create inappropriate—and perhaps dangerous—overconfidence.[338] And it is not clear such risks would be worth it anyways: using machines in this manner may turn out to waste, rather than save, time.[339]

But there is a clear desire for such a tool. And so, we should proceed cautiously. As a genie might say: "You ain't never had a friend like me."[340]

---

leaked out ahead of a decision"); Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1231–32 (11th Cir. 2024) (Newsom, J., concurring) (discussing potential for manipulation).

335. *See, e.g.*, Re & Solow-Niederman, *supra* note 57, at 246 ("AI adjudication's *development path* will affect not just how the technology is used, but also the legal system in which it operates."); Geddes, *supra* note 258, at 43–45, 50 (recounting significance of shifting "narrative power" as a result of machine use).

336. *See Ross*, 331 A.3d at 230–31 (Howard, J., concurring).

337. Crootof et al., *supra* note 25, at 468–69 ("[A] hybrid system can all too easily foster the worst of both worlds."); Jonathan H. Choi & Daniel Schwarcz, *AI Assistance in Legal Analysis: An Empirical Study*, 73 J. LEG. EDUC. 384, 397 (2024) (suggesting use of AI assistance in legal analysis may raise poor students' performance but lowers top students' performance); *see also* Solove & Matsumi, *supra* note 11, at 1935–37 (discussing "automation bias," the tendency to defer to machines and how "there might be significant pressures for humans to accept algorithmic output rather than deviate from it by using their discretion and emotion").

338. Kashmir Hill, *They Asked an A.I. Chatbot Questions. The Answers Sent Them Spiraling.*, N.Y. TIMES (June 13, 2025), https://www.nytimes.com/2025/06/13/technology/chatgpt-ai-chatbots-conspiracies.html.

339. *See* Joel Becker et al., Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity 2–3 (July 25, 2025) (unpublished manuscript) (on file with arXiv), https://arxiv.org/pdf/2507.09089 [https://perma.cc/VYM3-48QG] (early preprint finding "a large disconnect between perceived and actual AI impact on developer productivity" in sample of 16 developers completing 246 tasks); *cf.* Choi & Schwarcz, *supra* note 337, at 390, 397, 405–06, 409–11 (noting AI-assisted exam-takers exhibited variable improvements (or decreases) in quality correlated with task-type and baseline skill, but slightly improved speed).

340. ALADDIN (Disney 1992).

IV. CONCLUSION

One might have titled this paper the unreasonable effectiveness of uncertainty.[341] The problem of normative uncertainty has a certain structure. Analyzing it may subvert expectations, but in a way that I hope brings old thoughts into focus.

We began by recognizing a particular type of uncertainty, namely, normative uncertainty. It is the kind of uncertainty that a self-driving car can't help you resolve in a trolley problem: even if your self-driving car could perfectly execute whatever rule you chose, you don't know what the right rule is.

We did not try to solve this problem. Instead, we looked at what it might mean for the "right to a human decision debate." We engaged in a bit of necessary simplification, dividing the debate into humanists and machinists, even though everyone knows the question is contextual and even though everyone knows "it's humans all the way down" and sideways.[342] We did so in order to analyze the leading argumentative strategies for and against the idea that, in some contexts, there is a categorical imperative against using machines to make certain decisions.

Normative uncertainty revealed those arguments to be lacking.

The humanists' Arguments from Explanation had seemed increasingly strong as the machines grew increasingly—and necessarily—opaque. In contexts where a decision is owed, such machines cannot give an "aligned justification": one that connects the machine's actual decision procedure with the normative features that would justify its outputs. Humanists argue that this precludes replacing humans with machines in contexts like judging because machines cannot give you the right kind of explanation.

Unfortunately, normative uncertainty shows that human judges won't, either—at least, not if they respond rationally to normative uncertainty.

The machinist's Better Decision Argument turns the humanists' parade of principles into grist for the calibration mill: Fairness, accountability, transparency, dignity, autonomy—the lot—do not show that there is a right to a *human* decision. Rather, the question is whether machines or humans can do it better, and that depends on technical limitations of what the

---

341. An homage to two important papers in this space. *See generally* Eugene Wigner, *The Unreasonable Effectiveness of Mathematics in the Natural Sciences*, 13 COMMC'NS ON PURE & APPLIED MATHEMATICS 1 (1960) (examining how math has great explanatory powers for physics); Alon Halevy et al., *The Unreasonable Effectiveness of Data*, IEEE INTELLIGENT SYS., Mar.–Apr. 2009, at 8 (showing how effective data is for speech recognition and translation).

342. *See* Crootof et al., *supra* note 25, at 443 (collecting literature).

machine can do, or so the machinist's argument goes. Techno-optimists seem to think those limits non-existent, or at least, fewer than for humans.

Unfortunately (or fortunately?), normative uncertainty reveals the flaw with the machinist's Better Decision Argument: it is cold comfort to say you have a right to a "better" decision when you are unsure what "better" is, and standard model machines, by ignoring *normative* uncertainty, do not address the humanist's concern, but assume it away.

So, where does this leave us? Having torn down the leading arguments on both sides of the "debate," I suggested it was time to focus on a different question. Instead of asking who, when, and how with respect to human decisions, we should flip the script and ask when a machine might be an appropriate or an inappropriate tool. And the answer to that question turns, in part, on the right "scope of choice" for resolving normative uncertainty when engaged in the activity at hand.

I offered some suggestions for how we might think about answering that question, about which scope of choice is most suitable to which context. But I did not offer a methodology for answering.

Why? It is very hard. There might be disagreement, or uncertainty—or both!—about the appropriate scope of choice. For example, in the judicial context, one might think a narrower scope of choice—going issue by issue—is more appropriate.[343] Or one could choose a broader scope of choice, and frame the judge's choice as being about which jurisprudence—or school of jurisprudence—to follow at the start of their time on the bench.[344] Or you could do, as I have done here, and proceed ruling by ruling. There are considerations that cut in favor of each of these approaches and also against. The scope of choice is one of the most difficult aspects of normative uncertainty.

But despite uncertainty about the appropriate scope of choice, you can't proceed in the face of normative uncertainty without selecting one. And yet, that very selection can, paradoxically, affect the choice(s) ultimately made. For example, in the case of an appellate panel, the scope of choice selected for voting, be it issue-by-issue or case-by-case, is sometimes outcome-determinative because of how the issues sum to the case's resolution.[345]

---

343. LOCKHART, *supra* note 4, at 124–42; Cox, *Super-Dicta*, *supra* note 21, at 1596–99.

344. *Cf.* Richard Re, Essay, *Personal Precedent at the Supreme Court*, 136 HARV. L. REV. 824, 831–32 (2023) (arguing that judges do and should adhere to their previously expressed views); Cox, *Super-Dicta*, *supra* note 21, at 1596–99 (comparing scopes of choice).

345. *See* Kornhauser & Sager, *supra* note 41, at 11–12 (introducing the doctrinal paradox).

And so, as I noted at the outset, we can now see one of the reasons for the seeming intractability and discomfort within the human decision debate. It is not *merely* that machines crystalize questions of normative uncertainty and that the standard frame runs into rational errors. It is that the "right to a human decision" debate intersects with the problem of normative uncertainty *at the problem's most difficult joint.*

Alas, it may seem that in diagnosing the problem, I have merely replaced one set of hard questions with another. But diagnoses are also the first step in finding a path forward. And here I suggest we take a page from computer science.

Computer science has long been grappling with the value alignment problem: how do we specify an objective for the machine that won't backfire? And an important advance in computer science to address the value alignment problem is, counterintuitively, to recognize the uncertainty that the problem generates. That is, the new framework for AI has normative uncertainty—it is uncertain about what it ought to do. And so, it does the next best thing: it asks.

This is striking, but perhaps not surprising. Computer scientists often describe the value alignment problem in terms of wishes. They take the moral of these stories to be that it is difficult to specify what you wish for in a way that doesn't lead to unexpected results.

But there is also another moral: you should be careful what you wish for because you don't know what you actually desire—and even if you think you do, you might be wrong. Value alignment and normative uncertainty are friends.

I have attempted to provide language to distinguish them, so we can address them both appropriately. I hope this language helps lawmakers and regulators to better understand the problem that engineers face but struggle to articulate—and so to better evaluate and guide the solutions that engineers are devising. We need to understand when we are playing an assistance game with our machines, and what they are likely to interpret as our objectives. And if the machine says, "I don't know," sometimes we would be wise to say "I don't know" back.