Canadian Journal of Law & Jurisprudence XXXVII No.2 August 2024, 369-396

 doi: 10.1017/cjlj.2024.17

John Robots, Thurgood Martian, and the Syntax Monster:
A New Argument Against AI Judges

Amin Ebrahimi Afrouzi
Yale Law School, New Haven, Connecticut, USA
Email: a.e.afrouzi@yale.edu

Abstract

This paper argues that an AI judge is conceptually undesirable and not just something that lies beyond the state of the art in computer science. In a nutshell, even if an AI system could accurately predict how a good human judge would decide a particular case, its prediction would be the product of correlations between such factors as patterns of syntax in bodies of legal texts. This approach of AI systems is insufficient for basing their output on the sort of rationales that are expected of valid judicial decisions in any desirable legal system. Thus, by their very nature, AI systems are incapable of providing valid legal decisions in any such system.

Keywords: *AI judge; normativity; legal positivism; legal realism; intentional stance*

1. The Rise of AI Judges

Countless techno-optimists portend the arrival of AI judges that can someday purportedly substitute human judges.[1] There is a strong intuition against such a substitution, even if in some ideal future, the decisions of AI and human judges were indistinguishable.[2] In this paper, I defend this intuition by arguing that decisions made by AI judges are bound to be deficient qua legal decisions in at least

1. At this time, a google search for AI judges returns close to seventy one million hits (searching for that phrase in quotations returns more than twenty thousand). As just one example from the legal profession that's currently among the top hits, consider a post by an attorney from a NY litigation firm on the American Bar Association (ABA) blog, which claims: "it is hard to imagine a professional task better suited to AI that that of the American judiciary," adding optimistically that the arrival of AI judges is "only a matter of time." Christopher Michael Malikschmitt, "The Real Future of AI in Law: AI Judges" (18 October 2023), online (blog): *ABA* www.americanbar.org/groups/law_practice/resources/law-technology-today/2023/the-real-future-of-ai-in-law-ai-judges/ Morever, the idea of AI judges is sometimes expressed under alternative banners, such as *robo-judges*, *e-judges*, *lawbots*, and so forth.
2. For examples of others who share this intuition, see Ian Kerr & Carissima Mathen, "Chief Justice John Roberts is a Robot" (2014) University of Ottawa Centre for Law, Technology & Society Working Paper, online: papers.ssrn.com/sol3/papers.cfm?abstract_id=3395885; Kiel Brennan-Marquez & Stephen E Henderson, "Artificial Intelligence and Role-Reversible Judgement" (2019) 109:2 J Crim L & Criminology 137.

one crucial respect: they lack a normative rationale. This suggests that the task of human judges should not be delegated to AI.[3]

A significant advantage of my argument over the most common objections to AI judges is that it is not contingent on them making mistakes. Common objections to AI judges rest on the disparities between AI and human work products that can arise from differences in their respective capacities.[4] But with enough progress, AI and human work products may become indistinguishable. Thus, such common objections may no longer arise in the future. My argument, however, rests on an inherent feature of AI technology. As such, it imposes a general constraint on AI judges, irrespective of their advancement or particular embodiment, even if they someday succeed in producing work that is indistinguishable from humans'.

In a nutshell, my argument is that even if AI reaches the same outcome as human judges, it will do so on the basis of statistical correlations and causal inferences. Such reasoning, though in theory capable of predicting correct outcomes, will be necessarily deficient in rationale. Yet, we expect of judges of any desirable legal system that they not only arrive at the right outcome but do so on the basis of the right rationale.[5] This expectation will be necessarily unmet if AI judges were on the job, even if their decisions were indistinguishable from human decisions.

The objection to judicial automation that I shall articulate here is original, I hope, in its formulation and argumentative detail. But its motivating intuitions have been 'in the air' for a while.[6] Yet, both the academic and public discourses fail to appreciate its devastating impact on the quest for judicial automation. My main aim is to highlight the dooming nature of the argument by tightening some loose ends, repudiating persisting optimism, and presenting the issue in a maximally honest, accessible, and even-handed manner.

The structure of the paper will be as follows: I will first make an analogy between correlative and causal reasoning (section 2). To that end, I introduce

3. On the idea that this, and similar shortcomings, could be compensated by purported advantages of AI judges, see section 4.4.

4. See Tania Sourdin & Richard Cornes, "Do Judges Need to Be Human? The Implications of Technology for Responsive Judging" in Tania Sourdin & Archie Zariski, eds, *The Responsive Judge: International Perspectives* (Springer, 2018) 87, arguing that AI judges should not be used because, lacking empathy, they could not be as responsive to the emotions underlying the matters that come before them.

5. Much hangs on what this expectation amounts to. I discuss this in detail in section 4.1.

6. For instance, Mireille Hildebrandt's distinction between text-driven and data-driven law bears some similarities to my distinction between the syntactical stance and the intentional stance. See Mireille Hildebrandt, "Law as Computation in the Era of Artificial Legal Intelligence: Speaking Law to the Power of Statistics" (2018) 68:1 UTLJ 12. Similarly, discussions of the gap between textual prediction and legal normativity have appeared in various formulations in jurisprudence written in Civil Law jurisdictions. For English-language examples, see e.g. Mireille Hildebrandt, "Legal and Technological Normativity: more (and less) than twin sisters" (2008) 12:3 Techné 169 at §4.3; Emre Bayamlio lu & Ronald Leenes, "The 'rule of law' implications of data-driven decision-making: a techno-regulatory perspective" (2018) 10:2 Law, Innovation & Technology 295; Claudio Sarra, "Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of *Contestability by Design*" (2020) 20:3 Global Jurist 1.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

the *syntactical stance* as an explanatory strategy by drawing from and extending Daniel Dennett's discussion of the astrological and physical stances. The 'syntactical stance' is a stance in which the syntactical arrangement of a certain domain of texts features as the explanation for the arrangement of syntax in other texts that have been written in the past or the prediction of the arrangement of syntax in yet another group of texts that are to be written in the future. The syntactical stance is similar to the physical stance in that it relies on the constellation of certain phenomena to predict a future constellation of similar phenomena. The physical stance relies on the constellation of physical particles and forces to predict future physical events. The syntactical stance similarly relies on the constellation of syntax in existing texts in order to predict the constellation of syntax in future texts.

Next, I will introduce a thought experiment about deterministic universes, in which a Martian and a syntax monster predict—word for word—the decisions of a human judge from the physical and the syntactical stances, respectively (section 2.1). This means that the Martian considers the constellation of all matter and forces in the universe to predict the constellation of the document that the human judge would have written, had she been presiding. Similarly, the syntax monster considers the syntactical constellation of certain documents (including the parties' briefs and other documents they file) and then predicts the syntactical constellation of a document that a human judge would have written, had she been presiding.

I then argue that the syntax monster's work product will lack the requisite rationale, given it is made from the syntactical stance (section 2.2). To this end, I distinguish between citing a correlation as grounds for an explanation or prediction and citing normative reasons as the basis or the rationale of a legal judgment. I then argue that unlike the rationale underlying the decision of a human judge, correlations cited by the syntax monster are not responsive to normative inquiry, meaning that they could not even hope to justify the syntax monster's rulling.[7] That deficiency conflicts with a feature that we expect of legal decisions in any minimally desirable legal system, namely that they should not only be correct in their holding (i.e., what they mandate or prescribe) but also be correct in their rationale, wherein correctness of rationale requires at a minimum that it be responsive to normative inquiry.

Finally, I argue that barring outrageous assumptions, an AI judge will face the same problem that the syntax monster faces, because irrespective of its specific embodiments, its reasoning too will be correlative rather than normative (section 4). Obviously, the syntax monster takes inspiration from the operations of machine learning systems. However, my claim is not that machine learning, much less all forms of AI deliberation, are 'syntactical' in nature. Rather, my

7. I want to remain agnostic about whether the rationale behind the decision of the human judge actually justifies the decision's holding, though I assume that at least it *hopes* to do so. All I want to claim is that judgments from the syntactical stance lack such a rationale. I discuss the objection arising from the fallibility of human judgments in section 4.3.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

claim is that all forms of AI deliberation, including deliberation from machine learning, will have the same problem as syntactical deliberation because they too are correlative (hence unresponsive to normative inquiry). The thought experiment of the syntax monster is then merely meant to draw out the similarity between two modes of explanation (and prediction): causal and correlative. This is to show that correlative deliberation is as unresponsive to normative inquiry as causal deliberation. But my claim is not that machine learning is 'syntactical' and that moreover all AI technologies are like machine learning. It is rather that all AI technologies that embody correlative deliberation will fail to live up to the legal ideal in the same way that the syntax monster would. I proceed by responding to some objections and refute persisting arguments in favor of judicial automation (section 4). I then close with reflections on irresponsible techno-pessimism (section 5).

2. The Syntactical Stance

Dennett points out that for predicting the behavior of any system, various predictive strategies or "stances" can be adopted.[8] Some of these strategies will be more reliable than others. Taking the astrological stance to predict the prospects of a person, for example, entails feeding "the date and hour of the person's birth … into one or another astrological algorithm for generating predictions of the person's prospects."[9] The astrological stance likely does not work any better than chance, however.

A better predictive stance, Dennett offers, is "the physical strategy, or the *physical stance*."[10] Taking this stance to predict a system's behavior entails determining the system's "physical constitution (perhaps all the way down to the microphysical level) and the physical nature of the impingements upon it" and using the "knowledge of the laws of physics to predict the outcome for any input."[11] In a causally deterministic universe, taking the physical stance can at least in principle reliably predict phenomena of varying complexity, from "the effect of leaving the pot on the burner too long" to "the entire future of everything in the universe."[12]

In such a universe, what limits the availability of the physical stance in practice would be limits on our cognitive and sensory capacities. As Dennett writes: "The Laplacean omniscient physicist could predict the behavior of a computer—or of a live human body, assuming it to be ultimately governed by the laws of physics" from the physical stance.[13] There could be, in principle, "beings of vastly superior intelligence—from Mars, let us say" who could

8. Daniel C Dennett, "True Believers: The Intentional Strategy and Why It Works" in John Haugeland, ed, *Mind Design II: Philosophy, Psychology, and Artificial Intelligence* (MIT Press, 1997) 57 at 59ff.
9. *Ibid* at 59.
10. *Ibid* at 60 [emphasis in original].
11. *Ibid* .
12. *Ibid* .
13. *Ibid* at 66.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

"predict our behavior in all its detail" from the physical stance.[14] These Martians would be

> capable of comprehending the activity on Wall Street, for instance, at the micro-physical level. Where we see brokers and buildings and sell orders and bids, they see vast congeries of subatomic particles milling about—and they are such good physicists that they can predict days in advance what ink marks will appear each day on the paper tape labeled "Closing Dow Jones Industrial Average."[15]

Thus, in a deterministic universe governed by the laws of physics, the individual behaviors of all humans can, at least in principle, be predicted from the physical stance.

In the same spirit, I want to propose a stance that Dennett does not consider for the prediction of all texts that are to be written in the future. The stance entails analyzing the syntactical constellations of a vast body of previously written texts through rigorous statistical methods, calculating in detail the regularities with which certain patterns in syntax emerge, and finally, predicting the likelihood of the emergence of those patterns in future texts. Moreover, given variations in the tastes and styles of authors, in taking the syntactical stance, a predictor can put heavier weight on the patterns of syntax in the existing writings of a particular author for predicting her future writings, as compared to patterns of syntax in other writing. So too, it can consider the context in which the text is being written (email thread, novel, etc.), or its purpose (communication, entertainment, etc.). And so forth. Since this strategy relies on examining the constellation of syntax in existing texts in order to predict such constellations in future texts, I call it the *syntactical stance*.

2.1 Judging from the Syntactical Stance: A Thought Experiment

Consider now a syntax monster who is to the syntactical stance what a Martian is to the physical stance (the Martian is a Laplacean omniscient physicist). Just as the Martian inhabits a physically deterministic universe governed by causal laws, the syntax monster inhabits a syntactically deterministic universe governed by statistical correlations. That is to say that by hypothesis, the Martian and the syntax monster can predict the decisions of every human judge of their respective universes word for word and with one hundred percent accuracy. Now suppose that in their respective universes, people decide to delegate the jobs of human judges to these creatures instead. Consider now how these creatures would go about their job as judges.

The physical stance relies on the constellation of physical particles and the causal forces among them to predict future physical events. Thus, the Martian judge predicts the stable arrangement of ink particles on the paper (or keyboard strikes, light projections on monitors, etc.). The syntactical stance similarly relies on the constellation of syntax in existing texts and their statistical relations in

14. *Ibid* at 68.
15. *Ibid* .

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

order to predict the constellation of syntax in future texts. And so, the syntax monster judge predicts the constellation of syntax to be printed on the paper, projected on monitors, etc. The Martian judge explains each judicial decision, be it their own or otherwise, in terms of the arrangement of physical forces and particles elsewhere in the world. The syntax monster judge explains them in terms of the arrangement of syntax elsewhere in the world.

Note that *we* can describe the behaviors of the Martian and the syntax monster from various stances, including from what Dennett calls the "intentional stance."[16] Taking the intentional stance treats the entity whose behavior is to be predicted as a rational agent, ascribes to it a set of beliefs and desires (given its place in the world and its purpose) and, finally, predicts that this rational agent will act to further its desires in light of its beliefs.[17] But nothing of significance follows from this, because we cannot draw any conclusions about whether the entity in question actually is rational or whether it actually has beliefs and desires. As Dennett emphasizes, we can adopt the intentional stance to describe the behavior of a thermostat just as we describe the behavior of human beings. But this is no reason to think that either humans or thermostats literally have beliefs or desires. So too, the fact that the intentional stance can be adopted to describe the behaviors of the Martian and the syntax monster is no evidence about whether they have beliefs or desires.[18] I bring this up to emphasize that the reason we should conclude that the Martian and the syntax monster will judge from the physical and syntactical stances respectively is *not* the fact that *I* have chosen to explain their behaviors from those stances. The latter fact is no evidence for the former. Rather, I am *assuming* that this is how they go about their jobs *per hypothesis* and as part of the thought experiment's setup. This, and only this, is how they will go about their jobs because, for the purposes of a thought experiment, I have supposed that this and no more is what they are capable of. What is of interest to the discussion here is only the stance that *they* adopt, and in fact have no choice but to adopt, when *they* go about predicting (and explaining) the textual output of each court decision. What varieties of stances *we* could adopt to describe each of their respective behaviors is wholly irrelevant.

2.2 Lack of Rationale in Decisions

Suppose that the relevant legal outcome in a judicial decision is that Tom should pay Chris ten shillings. A problem arises when we inquire into the basis of that outcome. Consider the Martian judge first. If in response to our inquiry, he could

16. *Ibid* at 61.

17. "A little practical reasoning from the chosen set of beliefs and desires will in many—but not all—instances yield a decision about what the agent *ought* to do; that is what you predict the agent *will* do." *Ibid* [emphasis in original].

18. Just as the 'intentional stance' can be (and often is) adopted to describe AI behavior. See e.g. Amin Ebrahimi Afrouzi, "The Dawn of AI Philosophy" (20 November 2018), online (blog): *Blog of the APA*, blog.apaonline.org/2018/11/20/the-dawn-of-ai-philosophy/; John Zerilli, "Explaining Machine Learning Decisions" (2022) 89:1 Philosophy of Science 1.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

have explained his decision to us, he would have said something along the following lines: Tom should pay Chris ten shillings because such and such physical forces and particles were arranged in such and such a way at such and such point in time and so forth, such that a judgment to that effect would have been printed out of the court's printer, had humans been judging. This sort of explanation, however, is completely unresponsive to our inquiry.

Similarly, the syntax monster could only explain its decision on the basis that such pattern of English characters formed such and such constellations in such and such regularity in such and such sources of text and so forth. That sort of explanation too is entirely unresponsive to our inquiry.

The reason such explanations prove unresponsive is that our inquiry is not about what causes or correlations underpin the outcome but about its normative basis. A human judge could of course also offer a causal or correlative explanation. Social scientists offer this kind of explanation all the time. For instance, not having had lunch seems to cause or correlate with judges handing harsher sentences.[19] But this sort of answer would also be entirely unresponsive to our inquiry. What is more, if this sort of answer was all that a human judge offered, we would find it reasonable for Tom to leave with a chuckle and impunity. That is because, like a human judge, we understand that in this context, our inquiry is not a causal or correlative inquiry but a *normative* one.

What goes unmet in the cases of non-human judges is the expectation that the judge offers a correct *rationale* for her decision. We need not assume of course that this rationale actually justifies the decision. It may not. Nevertheless, it must be responsive to our normative inquiry. Even if it fails to actually justify the outcome, it must be at least in the business of justifying it; it must be the right sort of response. Meanwhile, the explanations that non-human judges offer for their decisions do not ring in the right register because they do not even hope to justify them.

Contrast our discussion now to two cases in which the same problem does not arise. First, suppose that instead of enlisting the syntactical monster to be a judge, we enlist them to become an economist and replace a human economist's daily market predictions. The monster's report can replace the actual report because the actual report does not need a normative explanation. The economist also bases her report on causal and correlative factors. Thus, the monster economist's output would not be deficient in the ways that a monster judge's outputs would be.

Second, suppose that instead, we task the monster with predicting what Anna, a literature critic, will write for tomorrow's newspaper about Jane's new novel.

19. See Shai Danziger, Jonathan Levav & Liora Avnaim-Pesso, "Extraneous Factors in Judicial Decisions" (2011) 108:17 Proceedings National Academy Sciences United States of America 6889. For a succinct summary of critical responses to Danziger et al., see Konstantin Chatziathanasiou "Beware the Lure of Narratives: 'Hungry Judges' Should Not Motivate the Use of 'Artificial Intelligence' in Law (2022) 23:4 German LJ 452-64. As Chatziathanasiou's title suggests, given live debate about the hungry judge effect's validity, we should revisit efforts to automate judicial decisions insofar as they are motivated by it. I address this specific conern on a different occasion. But note that my argument here does not in any way depend on the truth or falsity of Danziger et al.'s claims.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

The same sort of deficiency does not arise if Anna's editor decides to circumvent Anna and send the monster's output to print instead, because the rationale behind it does not matter. Where the monster fails is when the output's rationale is both of interest and expected to be responsive to normative inquiry.

3. Judging on the Basis of Correlative Reasons

We are now in a position to articulate exactly what relevant distinction we can draw between AI and human judges and exactly what legal ideal, if any, judicial automation would violate. In order to cast the strongest objection, let us assume the strongest case in favor of our AI judge. Let us then suppose that our AI judge can predict what a human judge in our universe would do (or would have done, had she been presiding). Let us further assume that this AI can make such prediction as accurately as the syntax monster or the Martian could do with respect to human judges in their respective universes. But we need not assume that our universe is deterministic in any similar way to their universes. All we need to assume is the same level of accuracy, however that is achieved.

Moreover, let us remain neutral to the exact specification of the AI in a way that can take into account any foreseeable advancement in computer science. This allows us to evaluate the ethics of deploying AI judges at a conceptual level.[20] Nonetheless, there is only so much that can be assumed in favor of potential AI judges of the future. Importantly, the assumption should not be so broad as to capture all the forms of non-human intelligence imagined in science fiction. Rather, even the AI judges of the future must be bound by—as they are enabled by—inherent features of our AI technology. It is this that distinguishes future AI judges from angels or omnipotent magical creatures. And we must be able to distinguish the two, because our thought experiment must be probative about *AI* and *its* ethics. So, the kinds of advances we are going to assume must be conceptually compatible with the sort of intelligence we are probing.

In order to capture all the *relevant* forms of AI, therefore, we can make the modest assumption (and the argument will be limited to this assumption) that our AI is empowered by a technology that roughly follows the trajectory of existing data-driven computational technologies that we constantly hear about these days.[21] Concretely, our assumption will be that our AI is a computational technology that generates its outputs (its decisions) on the basis of rigorous statistical correlations (and perhaps even on the basis of causal inferences, to the extent that's possible).

---

20. To my knowledge, Kerr and Mathen were the first to formulate the question in these terms: see Kerr & Mathen, *supra* note 2. For a more recent example, see Eugene Volokh, "Chief Justice Robots" (2019) 68:6 Duke LJ 1135 at 1191. For an example in the context of autonomous weapons, see Duncan Purves, Ryan Jenkins & Bradley J Strawser, "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons" (2015) 18:4 Ethical Theory & Moral Practice 851.

21. For the history of this trajectory, see Trevor Bench-Capon, "Thirty Years of *Artificial Intelligence and Law: Editor's Introduction*" (2022) 30:4 AI & L 475.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

With these assumptions, it is clear that AI judges (in our universe) will face the same problem that syntax monsters and Martians face in their respective universes. Because given these assumptions, we again know how they will go about their business and what the basis of their decisions will be. Granted, their decisions can be based on numerous other factors besides the arrangement of syntax in existing texts. But the *basis* of their decisions will nevertheless be correlations between patterns of those numerous other factors. And insofar as this much is true, the rationale of these judges will also be unresponsive to normative inquiries. This means that delegating the tasks of human judges in our world to AI would conflict with the legal ideal that judicial decisions be made on the basis of reasons that are responsive to normative inquiry.

Note that this critical point holds true of data-driven predictive computational technology generally and not just of machine learning or other limited specifications of AI.[22] The thought experiment of the syntax monster takes inspiration from the operations of machine learning systems. But the idea of the syntactical stance is only meant to bring out the similarity of correlative to causal reasoning and their difference from normative reasoning. As such, my criticism is not about systems that for decision-making *literally* rely on statistical correlations between syntactical patterns in existing texts. Nor is my claim that machine learning, much less all forms of AI deliberation, is 'syntactical' in nature. Rather, my claim is that all forms of AI deliberation, including deliberations of machine learning systems, will have the same problem as syntactical deliberation, because these too will necessarily be correlative (hence unresponsive to normative inquiry). This is because correlative reasoning is an inherent feature of all the *relevant* forms of AI—past, present, and future. Thus, the claim is not that machine learning is 'syntactical' and that moreover, all AI technologies are like it. It is rather that all AI technologies (that run on correlative reasoning) will fail to live up to the legal ideal in the same way that the syntax monster would.[23]

To sum up then, even if a future AI system could accurately predict how a good human judge would decide in a particular case, its prediction would be the product of correlative reasoning, such as reasoning on the basis of correlation

22. *Cf* John Tasioulas, who criticizes AI judges on different grounds (though in a similar spirit) but narrows his criticism specifically to machine learning: "let me pursue some worries about AI-powered adjudicatory tools under this heading in four steps, focussing on the deep-learning style of machine learning that has achieved prominence in recent years." John Tasioulas, "The Rule of Algorithm and the Rule of Law" in Christoph Bezemek, Michael Potacs & Alexander Somek, eds, *Vienna Lectures on Legal Philosophy: Vol 3: Legal Reasoning* (Bloomsbury, 2023) ch 2 at 14.

23. Obviously, futurologists can stipulate the existence of an artificially intelligent judge, whose decisions *do* have a normative rationale. What they would be stipulating, though, is a kind of artificial intelligence that is purely fictional, one that has nothing to do with computers, computation, or data. At this juncture in human history, talking about those kinds of human-made intelligence (which will surely be differently-branded from AI) is tantamount to talking about omnipotent magical creatures. We have no idea what these creatures or that kind of intelligent artifacts are or are not capable of. That is why I urge us to limit the discussions to *relevant* AI technology, which even when perfected, would rely on correlative reasoning, and hence, could not have a normative rationale. For why generative AI or other advancements in AI tasked with *normative* predictions don't change anything, see section 5.1.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

between such factors as patterns of syntax in bodies of legal texts. It is this approach of AI systems generally that is insufficient for basing their output on the sort of rationales that are expected of valid judicial decisions in any desirable legal system. Thus, by their very nature, AI systems are incapable of providing valid legal decisions in any desirable legal system.

Lacking the right rationale forbids the use of AI judges as well as thousands of existing laws and practices that, under the status quo, delegate similar normative judgments to AIs at various levels of the legal process. In many jurisdictions, there are already automated judgments, e.g., payment orders, where the judge simply copy-pastes into a judgment from a software. Any such practice will be implicated by my argument, meaning that if I am right, lawyers have every reason to object to them on behalf of their clients.[24] So too are *non*-legal decisions with normative upshots that are under similar (tacit or explicit, regulatory or otherwise) justificatory requirements (e.g., loan decisions). Such judgments can and should be challenged along the same lines, though I shall omit further discussion of them here.

4. Some Objections

In this section, I consider and respond to some objections and refute some persisting defenses of AI judges.

4.1 AI Outcomes Appear to Have Rationales to Us

In the thought experiment, I granted that AI judges can produce output that is in content indistinguishable from content created by human judges. This consists in every word that appears in regular judicial decisions, including what looks like the decision's rationale to human onlookers. Why is that not enough?

Eugene Volokh argues that "when AI judges become highly effective at crafting persuasive legal arguments, there will be little reason to prefer human judges to AI judges, at least for the overwhelming majority of legal questions, including the law development questions that reach the Supreme Court."[25] By "crafting persuasive legal arguments," Volokh does not mean that AI judges actually *base* their decisions on persuasive normative rationale. Rather, he means it in exactly the terms I have granted and consistent with our assumptions about AI judges: AI judges produce output that is in content indistinguishable from content created by human judges, including what looks like a rationale to onlookers. Indeed, Volokh adds: "I have carefully tried to avoid saying that the AI judge should explain why *it* reached the result it reached, saying instead that the AI judge should articulate reasons *supporting* the decision."[26] Volokh is careful to exclude the basis of AI's

24. For exceptions in which AI is merely deployed as a tool that features in law's mandate, see section 4.5.
25. Volokh, *supra* note 20 at 1191.
26. *Ibid* at 1164 [emphasis in original].

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

decision, presumably because he implicitly recognizes that such an explanation would be unresponsive to normative inquiry. His suggestion is therefore that the existence of supporting reasons for legal judgments is sufficient for them to be valid. However, although the existence of supporting reasons is likely necessary for the validity of legal judgments, it is likely not sufficient for it in any desirable legal system.

Obviously, according to our assumption, the AI presents the reasoning supporting its judgment *as* its rationale in exactly the same manner that a human judge presents hers. Thus, to onlookers, this part of the output appears as though it really were the AI's rationale. Without knowing the assumptions we have made earlier, there would be no way of telling the difference. It is only thanks to what we know to be true generally about the sort of AI technologies we are considering, and in light of what our thought experiment shows about correlative and causal inferences, that we can be sure that what the AI presents to us is not in fact the basis of its decision. In fact, for all that observers can see, the rationale of the human judge may be no more than the same. But we are certain about the rationale of the AI judge what we can only suspect about that of human judges.[27]

Once we recognize this fact, it becomes evident that the supporting reasons, which the AI presents as its rationale, are in fact no such thing and therefore quite irrelevant to meeting the legal ideal of our concern. This is because we know that—contrary to appearance—such content would have had nothing to do with the judgment's actual basis. The judgment's actual basis, as Volokh seems to acknowledge, was the correlation of various factors that the AI considered, just as the arrangement of stars and planets was the actual basis of the Martian's decisions.

So Volokh's appeal to the sufficiency of the output is an appeal to the sufficiency of the decision's *ex post* justifiability. But this is surely not enough. For consider that even bare-bone AI judgments may be justifiable *ex post*, say, by onlookers. The decision in both cases, though justifiable *ex post*, lacks a normative basis, and thereby fails to meet the relevant legal ideal. That the *ex post* justification is produced by the AI rather than onlookers does not change anything.

But Volokh's stance may be that *ex post* justification is all we need insofar as *it* has a hope of justifying the outcome. For after all, *it* is responsive to our normative inquiry "why." But requiring a rationale for a decision is not simply requiring that there be *some* content in the world with a hope of justifying it. It is rather that *the basis on which that decision was made* has had a hope of justifying it. Obviously, there may be *ex post* justifications even for the decisions of the Martian and the syntax monster judges. Nevertheless, it seems plausible that Tom can still walk away with a chuckle after pointing out the obvious fact that the judgment of these creatures, though justifiable, was made on the basis of things like the arrangements of planets or syntax. Presumably, Tom can point to

27. I discuss the normative upshots of this suspicion in section 4.3, below.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

this mere fact and walk away with a chuckle, because considerations such as the arrangements of planets and syntax are not valid bases for legal decisions.

Finally, perhaps Volokh is contesting this last point. Perhaps he is inviting us to consider a system of judging in which decisions made on the basis of the arrangements of planets *are* valid, so long as they are accompanied by *ex post* justificatory content. Put another way, Volokh may be inviting us to jettison the ideal for legal judgments to have a normative basis, if we ever become successful in designing a system that can produce justifiable outcomes via other means. We must take this proposal extremely seriously, given the extraordinary benefits that per hypothesis flow from effortless, instantaneous, and accurate judicial decisions by Martian, syntax monster, or AI judges. Because it is precisely in light of such extraordinary benefits that Volokh suggests that even decisions made on the basis of the arrangements of planets should be valid, so long as they are accompanied by persuasive *ex post* justificatory content. This proposal seems quite reasonable because the qualification made by justification is quite strong, which is something we have granted by stipulating that the content of the decisions of these imaginary judges will be identical to human decisions. This stipulation entails that their decisions will be equally *ex post* justifiable.

Volokh's view, so interpreted, may strike some, especially non-American readers, as an extreme Holmesian line of thought.[28] This Holmesian stance is widely suspected to be implausible as a model of the concept of law.[29] But its conceptual (im)plausibility should not concern us on this occasion.[30] What should concern us rather is a purely normative question: whether or not we should adopt such an institution of judging (be it in accordance with the concept of law or the concept of some other system of social control) in pursuit of the promised benefits of judicial automation. The correct answer to this purely normative question, I argue, is in the negative.

My reason is that jettisoning the ideal that decisions have a normative basis makes for an objectionable institution of judging, even when the decisions of such an institution are *ex post* justifiable. And this is all the grounds we need to reject it. Consider that the rationale judges offer under the status quo may actually fail to fully justify their decisions. The full justification of decisions under the status quo, if they can indeed be fully justified, is *institutional* and rests in part on the *ex ante* legitimacy of the process that includes, among others, a constraint that judicial outcomes be reached on the basis of a normative rationale rather than syntactical correlations or physical causes. But by rejecting this constraint, the Holmesian view cannot avail itself of the institutional justification that could

28. "The prophecies of what the courts will do in fact . . . are what I mean by the law." Oliver W Holmes, "The Path of the Law" (1897) 10:8 Harv L Rev 457 at 461.

29. See HLA Hart, *The Concept of Law*, 2nd ed (Oxford University Press, 1961) at 1-2.

30. The reason I flag this matter, though, is that some readers familiar with Hart's critique of the Holmesian view may wonder whether what I am about to say intersects with that debate. It does not. My point in this section is normative, so it is neutral about who wins the Hart-Holmes debate. A legal system of the kind Holmes has in mind may be (*pace* Hart) conceptually possible. Our question is whether we would want to adopt such a legal system as our own.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

(at least in theory) close the gap between the human judge's rationale of her decision and its actual justification. It is not clear moreover, whether and how, an institution of judging on the basis of the arrangement of planets can make up for the points it loses in the ability to claim legitimacy over cases where the supplemented justificatory content fails to actually justify the decision *ex post facto*.

Indeed, this problem runs deeper and affects even cases where the supplemented justificatory content does happen to actually justify the decision *ex post facto*. Consider that in many legal cases, we just do not know what the *right* outcome is, even if several outcomes would be *ex post facto* justifiable (or conversely, if none would be). In other words, there may be no *right* outcome *ex ante*. We would then need a legal process to settle what outcome is right. The rightness of the outcome, in other words, will in part rest on it being the outcome of processes that can be defended against alternative processes *ex ante*.

Existing practice under the status quo can be (at least somewhat) defended on grounds that its process decides cases on the basis of reasons. To the extent this defense succeeds, the outcomes of *this* process can be regarded as right. But an institution of judging on the basis of the Holmesian view could not be *ex ante* defended, at least not as an alternative to the status quo, because its process decides cases on the basis of causes and correlations. So, its outcomes cannot be regarded as right in the same sense (at least not when the status quo process is more choice-worthy), even in cases where supplemented justificatory contents do happen to justify these outcomes.

We must therefore reject a Holmesian institution of judging, that is, we must reject an institution of judging that regards the arrangements of planets and syntax as valid bases to decide legal cases, even if such an institution is qualified by the requirement that the resulting decisions be justifiable *ex post facto*.

At this juncture, Volokh may try to qualify his stance by maintaining that we should only treat *those* decisions of AI judges as law that are *actually* independently justified *ex post facto* to the parties involved. Perhaps if the outcomes are actually justified to the parties, albeit *ex post facto*, their shortcomings will be overcome. Perhaps. Even so, the doors to such a qualification are closed. Because sincerely justifying each outcome would require someone like a human judge to review AI decisions *de novo*. Barring practical issues, this may compensate for the lack of rationale in the initial AI-generated decisions, but it also renders them entirely redundant.

Note that the door to limiting *de novo* reviews to only those decisions that are appealed is also closed. For there is no reason to think any party who receives adverse decisions by an AI judge would not appeal on the grounds that the decisions lack rationale. Indeed, one could regard this very paper as a class action appeal on their behalf.

Finally, practically speaking, a right to appeal and *de novo* review may not in reality fully protect the affected parties, because *de novo* review may be biased in favor of affirming or overturning AI-generated decisions. For one, human judges, depending on their attitude to technology, may show more or less deference to AI-generated decisions than to decisions made by a mortal judge down the

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

hallway. But they may also be too languid, over-worked, or apathetic to review the matter sincerely *de novo*. Practically speaking, the temptation to simply affirm on the basis of the justificatory content AI generates may be too hard to resist.

4.2 The Argument Rests on Controversial Assumptions

4.2.1 Legal Positivism

I argued that judicial decisions must have a rationale with the hope of justifying the holding in normative rather than correlative or causal terms. To begin with, it may be tempting to think that in maintaining it, I have begged the question against legal positivists. Legal positivists maintain that law need not be morally justifiable in order to be valid.[31] In arguing that judicial decisions need a rationale to be valid, it is tempting to think, I have assumed legal anti-positivism.

Joshua Davis for instance argues that AI judges would only be unfit if, *á la* anti-positivists, legal interpretation partly consisted in deciding what the law *should* be, from the standpoint of moral justifiability.[32] If, however, *á la* legal positivists, legal interpretation entirely consisted in describing what the law is, or how it would be interpreted by others, then AI judges would be as fit for the job as human judges.[33] If he were right, whether AI judges can replace humans would have depended on the right answer to a central question in general jurisprudence. Then, my case against AI judges would fail because of illicitly assuming anti-positivism.

But as I argue in a moment, my account is entirely consistent with legal positivism. In fact, my case against AI judges circumvents all debates in general jurisprudence. This in turn shows that whether or not we should adopt AI judges has nothing to do with the right answer to the positivism/anti-positivism debate.

Legal positivism maintains that moral justifiability of legal decisions is not necessary for their legal validity.[34] This is something I can easily grant. As discussed in the previous section, the requirement that outcomes have a rationale is different from them being morally justifiable. And as should be clear from my discussion there, my insistence that outcomes have a rationale is completely consistent with them actually failing to be justifiable.

True, I do maintain that legal decisions must be made on the basis of criteria that hope to morally justify them in order to be legally valid. But this is neither a moral constraint on legality nor a conceptual claim about 'the law' as such. It is rather a *legal* constraint on legality that *contingently* holds in most existing legal systems and for good reasons. Let me unpack these issues in turn.

---

31. See Hart, *supra* note 29 at 207-12.

32. See Joshua P Davis, "Law Without Mind: AI, Ethics, and Jurisprudence" (2018) 55:1 Cal WL Rev 165, advocating against AI judges, arguing *á la* anti-positivists that sometimes judges look to law as a source of moral instruction. This, however, requires the ability to reason morally. But given that AI cannot (yet) reason morally, AI could not fulfill the function of human judges.

33. See *ibid*.

34. See Hart, *supra* note 29 at 207-12.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

Consider first, that on my account, the criterion that legal outcomes be correct in their rationale is a *legal* rather than a moral constraint on legal validity.[35] The question I address is not whether we should treat AI-generated decisions as *morally* valid. It is whether we should treat them as *legally* valid. My answer to the second question is that we should not—because *legally valid* decisions must be correct not only in their holding but also in their rationale, and yet, AI decisions cannot even hope to be correct in their rationale.

Second, on my account, even this constraint is not a necessary constraint for either the concept or the nature of law. In arguing against the Holmesian view above, I readily granted that a legal system that makes decisions without rationale is conceptually possible but argued that such a *legal* system would be objectionable. My case against AI judges is therefore neutral about what law per se is or is not. It simply says that 1) we expect of judges to decide cases on the basis of normative rather than purely causal or correlative basis; and 2) it is a good thing for us to continue to expect this of judges, even if not doing so is conceptually possible.

4.2.2 The Chinese Room Argument

My talk of syntax and syntax monsters may give some readers the wrong impression that I claim (or worse, that I presuppose) that computation is necessarily syntactical or that machines necessarily lack semantic understanding or some such. The Chinese Room Argument famously maintains that a translation software does not understand the sentences it translates, even if it passes the Turing test.[36] But whatever its strengths and flaws, this argument is completely irrelevant to mine. As far as I am concerned, we can even assume that machines, syntax monsters, and Martians fully understand the semantics of their judgments. All the same, they will face the problem I have stipulated.[37]

4.2.3 Machines Are Unconscious, Immoral, or Inscrutable

My argument also does not assume that computers are unconscious in order to claim that they are incapable of moral reasoning.[38] Instead, it assumes that the

35. I take this point to be uncontroversial as a matter of existing practice. Consider, for instance, countless cases that are reversed and remanded on precisely these grounds, as well as appellate cases with concurring (or dissenting) opinions.

36. See David Cole, "The Chinese Room Argument" in Edward N Zolta, ed, *The Stanford Encyclopedia of Philosophy*, online: plato.stanford.edu/entries/chinese-room/.

37. Relatedly, some may think I have begged the question against functionalism. But I have not. As far as I am concerned, functionalism may be also true. My argument is not that a difference remains between two systems that are functionally identical. It is rather that the two systems are not functionally identical. To put it crudely, he says, it walks and talks like a duck but it is not a duck. I say it may talk like a duck, but it certainly does not walk the walk. My argument is therefore consistent with functionalism being true.

38. One of the main debates in the ethics of AI is about 'artificial moral agents' (AMA's). See e.g. Dorna Behdadi & Christian Munthe, "A Normative Approach to Artificial Moral Agency" (2020) 30:2 Minds & Machines 195. This debate is roughly about whether artificial entities,

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

relevant kind of AI technologies will make their decisions on the basis of correlations, regardless of how they are improved in the future. This is a modest assumption and entirely compatible with the possibility that AI systems may become conscious. All I claim is that these systems—whether or not aptly described by labels like 'intelligent', 'conscious', or 'moral'—could not respond to normative inquiry about the basis of their decisions.

Nor do I deny that computers can engage in moral reasoning. Technologists and ethicists have long strived to imbue artificial minds with moral values.[39] Recent scholarship even considers simulated value pluralism,[40] moral disagreement, and moral uncertainty[41] in machines. I leave it up to the reader to decide how much of this is actual, fictional, or rhetorical in existing embodiments.[42] What is important to my discussion is that these efforts are strictly to align machine behavior, including its 'reasoning' behavior, with morally justifiable behavior. Thus, it is an effort to produce autonomous cars that choose the lesser of two evils, autonomous assault drones that only fight 'just' wars—if there is such a thing, and AI judges that only convict the guilty.

I can grant the success of all such innovations in the future, because all such efforts aim to amend machine *behavior*. The point I am trying to make is that no concomitant efforts can be made or even conceived to empower the sort of machines we are probing with the capacity to contemplate the right kinds of *rationale*. Thus, even if, say, an autonomous human resources manager can be taught to select applicants based on merit rather than race, it could not ever tell us *why* it selects the people it selects, in a way that would be responsive to a normative

like computers and robots, are able to do wrong or be considered responsible for such wrongdoing. One controversial issue in this debate is whether artificial entities could be considered moral agents without having a conscious mind. As the main discussion should make clear, the upshot of this debate is irrelevant to my argument, which directly offers a negative response to the question whether we should task AIs with imposing normative demands on people. Indeed, upshots of the AMA debate may not be relevant to normative inquiries about whether AI should be tasked with decisions with moral upshots, let alone those that require a fitting rationale as well. Behdadi and Munthe write that the AMA debate "should focus on how and to what extent such entities should be included in human practices normally assuming moral agency and responsibility of participants" while observing that "this can be done without ever using the term 'moral agent', thus avoiding much of the conceptual confusions that have confounded a lot of the AMA debate so far" (*ibid* at 195, 212).

39. For a succinct overview as a well as a classic intervention in this domain, see Jan Gogoll & Julian F Müller, "Autonomous Cars: In Favor of a Mandatory Ethics Setting" (2017) 23:3 Science & Engineering Ethics 681.

40. See Jake B Telkamp & Marc H Anderson, "The Implications of Diverse Human Moral Foundations for Assessing the Ethicality of Artificial Intelligence" (2022) 178:4 J Business Ethics 961.

41. See Kyle Bogosian, "Implementation of Moral Uncertainty in Intelligent Machines" (2017) 27:4 Minds & Machines 591; Andreia Martinho, Maarten Kroesen & Caspar Chorus, "Computer Says I Don't Know: An Empirical Approach to Capture Moral Uncertainty in Artificial Intelligence" (2021) 31:2 Minds & Machines 215.

42. See Ibo van de Poel, "Embedding Values in Artificial Intelligence (AI) Systems" (2020) 30:3 Minds & Machines 385. For critical discussion, see Amitai Etzioni & Oren Etzioni, "Incorporating Ethics into Artificial Intelligence" (2017) 21:4 J Ethics 403, arguing that so long as AI-equipped machines comply with the law and the individual choices of their owners, a significant part of the ethical challenge posed by them can be addressed without the machines ever having to make moral decisions themselves.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

inquiry. The thrust of my argument is that even when machine behavior is extensionally indistinguishable from morally justifiable human behavior—that is, when the AI selects exactly the same applicants as unbiased humans would—it still could not tell us *why* it chose those applicants any more than 'what correlates with what' from something like the syntactical stance.[43] As I have emphasized earlier, this is not a limitation on using AI in all contexts, but it is a strong limitation on using it in the context of AI judges because judicial decisions must be made on the basis of normative reasons.[44]

Finally, one may object that I have illicitly assumed that AI reasoning will forever remain inscrutable to us. To see why I have done no such thing, it may be helpful to juxtapose my argument with so-called "black box" arguments against AI judges in the literature.[45] Black box arguments maintain that neural networks in machine learning are like black boxes, i.e., there is no understandable account of their operation. Accordingly, they worry that as a result, AI may malfunction in ways that are either unpredictable or inconspicuous. One worry here is that because we cannot always make sense of how AI yields the outputs it yields, we may not be able to predict with certainty that it would do the right thing (or refrain from doing something unintended and/or immoral). For instance, the worry may be that AI finds correlations inconspicuous to us and bases distributive decisions on them that amount to wrongful discrimination—in a way that may be not only unintended but also hard to detect or prove.[46] I however do not assume that AI decisions are necessarily inscrutable or unexplainable.[47] Nor is my argument that AI reasoning could possibly lead to unforeseen or incorrect decisions. My argument is rather that AI reasoning would be deficient in rationale even if it only led to foreseen or correct decisions and even if the basis of its decisions were readily accessible to us.

---

43. Even if its decisions are extensionally indistinguishable from that of humans, even if the content of its 'moral reasoning' is indistinguishable from that of humans, in short, however many times we pass the ball, AI rationale ultimately boils down to correlative predictions from something like the syntactical stance. Note that creating AI systems to detect rule violations in a way that is extensionally indistinguishable from that of humans may not be all that futuristic. For discussion, see section 5.1 below.

44. Whether automated hiring decisions are subject to such a constraint is a live debate, which lies outside the scope of this paper. See Kathleen Creel & Deborah Hellman, "The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems" (2022) 52: 1 Can J Philosophy 26, arguing that decisions in contexts such as automated hiring are not subject to such a constraint.

45. For a recent discussion of black box arguments in this context, see Ashley Deeks, "The Judicial Demand For Explainable Artificial Intelligence" (2019) 119:7 Colum L Rev Symposium: Common Law for the Age of AI 1829.

46. There are other worries in this area and further worries unique to the context of adversarial trials: see Tasioulas, *supra* note 22.

47. For ongoing research to address such objections, see Carlos Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence" (2021) 34:2 Philosophy & Technology 265. For challenges to the validity of this research, at least in a legal context, see Sebastian Bordt et al, "Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts" (2022) FAccT '22: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency 891, online: dl.acm.org/doi/10.1145/3531146.3533153.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

4.2.4 The Mind-Body Problem

Finally, have I not assumed that *human* reasoning is not syntactical, or correlative, or causal at some deeper level? And have I not assumed that human intellect involves a separate, non-physicalist or non-material mind, or some such, thereby begging the question in favor of dualism or against physicalism, thereby disregarding controversies in the so-called mind-body problem? No. Nothing I have said makes any assumptions about how human brains work or do not work.

4.3 Human Judges Also Do Not Know the Reasons Behind Their Decisions

It may be tempting now to think that where AI judges fail, human judges fail too. I have argued that AI judges ultimately make decisions based on statistical correlations which bear no relevance to normativity. It may be argued that human judges also ultimately make decisions based on biological or neurophysiological processes that bear no relevance to normativity.

To begin with, this argument proves too much. Showing that humans are also incapable of making legally valid demands on one another is no salvation for replacing them with AI judges. Instead, it undermines the legal process in both its human and machine-operated forms. As such, it is not something I need to worry about in my critique of AI judges. Nonetheless, I think the argument is mistaken too.

First, this objection rests on a defunct analogy between neurophysiological processes of the brain and the statistical processes of an algorithm. First, consider how this analogy conflates two levels of descriptions: if the analogy is supposed to operate at lower-level processes that give rise to computation, then the analogue to neurophysiological processes of the brain would be the electrical activity in a computer's circuits. The statistical processes in algorithmic decision-making are instead themselves the *higher-level* or emergent phenomenon of such electrical activity. The analogues to statistical processes are not the neurophysiological processes of the brain but the emergent phenomenon they give rise to, namely normative reasoning (and other kinds of *mental* phenomena).

Second, even if the analogy were apt in its level of description, it flaunts the widespread consensus that mental phenomena such as normative reasoning are not reducible to the neurophysiological phenomena they supervene on, unlike algorithmic processes that are reducible to the electronic processes occurring in their circuits.[48] The objection urges that if normative justification cannot obtain by way of electrical activity or algorithmic calculations, it could also not obtain

48. This is indeed one of the upshots of Dennett's argument discussed above; see Dennett, *supra* note 8. One of Dennett's points is that although the Martian can discern every human activity at the neurophysiological level, it cannot discern and would have no conception of mental phenomena like beliefs, hopes, or desires. Whatever the Martian cannot hope to observe is not reducible to its lower-level causes. For discussion, see John R Searle, *Mind: A Brief Introduction* (Oxford University Press, 2004) at 84-95. No need to say that on the other hand, the Martian discerns the algorithmic decision-making of computers down to their every detail. For where Dennett's analysis goes wrong, see Willem A Labuschagne & J Heidema, "Natural and Artificial Cognition: On the Proper Place of Reason" (2005) 24:2 South African J Philosophy 137.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

by way of biological or neurophysiological activity. That point is of course true. But it is irrelevant because I have not claimed that the normative reasoning which human judges are capable of obtains at the biological or neurophysiological levels. Instead, the idea here is that this sort of justification happens at the level of conscious human reasoning. For the objection to cast any doubt on the matter, it should either show that normative reasoning cannot obtain by way of conscious human reasoning, or that conscious human reasoning is reducible to the neurophysiological phenomena it supervenes on.

Alternatively, this objection can be cashed out as a worry that human reasoning is opaque to us, and perhaps even to the judge herself. The claim here may be that 'black box' arguments (which we met earlier) also apply to human reasoning. The worry would then be that human judges could act for the wrong reasons behind our backs, even unbeknownst to themselves. For instance, they could make distribution decisions as a result of implicit bias. Worse yet, the claim against human judges may be even stronger, in the same way that my critique of AI judges is stronger than black box arguments. Here, the worry would be that even when human judges decide correctly, their rationale is unknown, unknowable, or nonexistent, even though what they write down as their rationale looks like conscious legal reasoning to onlookers. In other words, the objection may be that what I have claimed of AI judges may be true of human judges as well.

This iteration of the objection again proves too much (showing that human judgment also lacks rationale is no salvation for AI judges). Nonetheless, I think its assumptions about human limitations are mistaken. It is of course important for us as humans to be realistic about the extent of our own capacity in normative thinking and about the limitations of the rationales of the demands we make on one another, sometimes as judges and under the banner of law. A grain of humility should make us pause, especially when the burdens and consequences of such demands are significant and grave. This level of humility and the skepticism that underlines this iteration of our objection, may ground a strong case, say, against the death penalty. But this kind of humility should not lead us to give up on institutional normativity altogether. For even if we fail to make the right choices in all the close calls, human experience attests that we can easily distinguish between significantly better or worse outcomes not just in hindsight but also *ex ante*.

Finally, this objection may be directed at juries rather than judges. Thus far, I have only discussed judges making decisions. But in many cases, including more serious cases, the verdict is decided not by a judge but by a jury. And one thing a jury cannot do is explain its rationale. After all, jury members vote and may have different reasons for their vote. And in any case, their deliberations are kept secret. A plausible iteration of our objection then may be that jury verdicts also lack rationales. To my lights, this is the only objection we may need to continue to worry about, because it is powerful. But several avenues of addressing it come to mind, which I shall briefly sketch without fully defending.

The first avenue would be to appeal to the distinction between questions of fact and questions of law. The idea here is that juries are determinants of fact

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

rather than determinants of questions of law. The response would then rely on this distinction and maintain that only answers to questions of law require normative rationales. After all, causal or correlational explanations *are* apt for determining matters of *fact*. The idea here is that what jury members do is like what the economist of my earlier discussion does (see section 2.2). But if this is right, then we will have to be ready to also maintain that the AI-jury-member's vote would not be deficient in the same way that the AI-economist's output was not. This may not be a terrible outcome (see section 4.5, below).

The second avenue of response would be to insist that the decision of each jury member, and hence each vote, has a normative rationale. Thus, each vote meets the requirement that it be cast on the basis of a normative rationale, though not being accompanied by it. There obviously still remains a gap between each *vote* having a rationale and a jury verdict as a whole having one. This gap could be bridged either by loosening the constraint of what we expect of valid legal decisions or fictionally ascribing a rationale to the collective verdict (along similar terms that we ascribe responsibility or agency to, say, corporate entities).

The third avenue would be to distinguish the *ex post facto* rationalization of the jury verdict (by the judge) from the *ex post facto* rationalization of AI. The distinction here will have to rest on the fact that jury members do have a normative basis for their vote, albeit one that is secret and perhaps differs from the rationale of other members, and so forth.

I am not myself entirely sure that these avenues, if pursued, will yield a fully satisfactory response to the objection's iteration against jury verdicts. But they might. And there may be other avenues. Else, we may have to qualify the scope of the case against AI judges to non-jury trials.

4.4 Access to Justice

What about the purported advantages of AI judges? Can they offset the lack of rationale in their decisions? The purported gains will be in accuracy, efficiency, or both. It is no secret that court dockets are full, the legal process is slow, humans are fallible (their decisions are prone to errors as a result of things like not having had breakfast), there is significant inconsistency in human-generated results, and so forth.[49] It is also no secret that the poor and marginalized get the brunt of all this. But these shortcomings arise from underfunding and understaffing the courts, which is a political *choice*. And as Wilmot-Smith forcefully argues, there is much that we can and should do to address the shortcomings that does not come at the cost of giving up any social, moral, or legal ideals.[50] The impetus behind Volokh's intuitions and the techno-optimist enthusiasm about judicial automation, therefore, is a false dichotomy, as if we had no choice but between

49. See Daniel Kahneman, Olivier Sibony & Cass R Sunstein, *Noise: A Flaw in Human Judgment* (Little, Brown Spark, 2021). See also Cass R Sunstein, "Governing by Algorithm? No Noise and (Potentially) Less Bias" (2022) 71:6 Duke LJ, 1175.

50. See Frederick Wilmot-Smith, *Equal Justice: Fair Legal Systems in an Unfair World* (Harvard University Press, 2019).

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

backlogs and AI judges. The hard question, however, is not about picking between backlogs and AI Judges. It is rather about the price we are willing to put on normative rationales for judicial decisions.

4.5 Breathalyzers and Traffic lights

It might be thought that a host of electronic devices like traffic lights and breathalyzers already do what an AI judge would do, albeit being much simpler and making much simpler decisions. If my argument outlawed such banal legal instruments, it would be overinclusive. Conversely, it may appear that such instruments amount to counterexamples to my analysis: if traffic lights can decide who should stop or go, then an AI judge could decide who should go to jail or walk free. But such objections fail, because they rest on a mistaken analogy between instruments like traffic lights and AI judges. Importantly, unlike AI judges, traffic lights and breathalyzers do not make any legal demands but merely feature in demands made by the law. The objection appears tempting if and when we think of the role of a stop light as deciding who should stop or go, or as demanding that we stop when it is red. But traffic lights make no such decisions or demands. It is rather the law that make such demands as 'stop when the light is red'. Thus, traffic lights and other such instruments merely feature in what the law demands.

5. Irresponsible Techno-Pessimism

There are two common mistakes in critiques of various uses of AI technology and existing critiques of AI judges are no exception.[51] The first mistake denigrates AI technology to claim that it falls short of the standards we hold humans to, when it does not. The second mistake pretends we hold humans to higher standards than we actually do and then show that AI falls short of those higher standards—in effect, holding AI to higher standards than humans. Arguably, such mistakes not only amount to bad philosophy but also create ethically risky policy discourse. This is because automation may come with great benefits, which means that halting it, where we should not, may come at great moral loss. Thus, just like unfounded techno-optimism, unfounded techno-pessimism is ethically irresponsible. Given that I have been critical of an equally common kind of irresponsible techno-optimism, one that elevates AI technology beyond bounds to claim that it

51. The classic piece in this literature is Kerr & Mathen, *supra* note 2, who argue that AI judges are incapable of judging because the capacity to judge is tied to membership of the community, and AI judges are not members of the community over which they preside. Obviously, though, membership in a community is not a constraint on the *capacity* to judge (think colonial courts and judges). Nor am I convinced that *membership in a community* is a standard to which we should necessarily hold human judges. But even if *membership in a community* were such a standard, it is not clear that AI judges fail to meet it. For further discussion, see Amin Ebrahimi Afrouzi "Role-Reversible Judgments and Related Democratic Objections to AI Judges" (2023) 114 J Crim L & Criminology Online 23.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

meets (or surpasses) the standards to which we hold humans, it is appropriate to discuss how I steer clear of said forms of irresponsible techno-pessimism before closing.[52]

5.1 Not Denigrating AI Technology

Starting from my characterization of AI technology and its future. Readers may find themselves in agreement with my rejection of the futurologist claim that 'in some future, AI can be capable of anything', but nevertheless worry that in some future, AI systems will be capable of reasoning in a way that is responsive to normative inquiry. So, at the risk of somewhat repeating what I have said above in section 4.2.3, let me explain in more detail why the door to such a future is closed.

All AI technology (of the relevant computational type) stems from assuming a form of determinism about the universe, including human behavior. The assumption, call it *modest correlative determinism*, is that all future events, including all future human behavior, are at least to an extent a correlative function of some past events. It then discerns (or 'learns from') the events in the past to make predictions about the future. To the extent that the universe and human behavior are not deterministic (i.e., to the extent they deviate from past correlative patterns), AI will be out of luck. But to the extent they are deterministic, they will be (in principle) predictable by AI, given sufficient relevant data, computational power, and advancements in the technology.

In practice, AI does not have access to all the correlative data. Thus, AI's predictive (in)accuracy will be a function of the training data and often falls short of what we can call 'the ideally attainable level of accuracy'. The challenge to reach the ideally attainable accuracy arises either from not *knowing* exactly what data about which past will be sufficient input for the AI to reliably predict future events, or not *having* it. Having is a matter of data collection. But more is not always better. What AI needs is enough of the relevant data. If you have AI predict consumer shopping behavior on the basis of astrological data, the predictions will not be good because astrological data are not strong predictors of human shopping behavior. A better predictor may be factors like their shopping history, income bracket, geographical location, height, weight, or some combination of several such factors. The *relevant* data is the right combination of factors that amounts to the best possible prediction. When we talk of the ideal AI technology of the future, we are talking about AI systems that in addition to having sufficient computing power, have all the relevant data for our target prediction domain, such that they can predict events in that domain at the ideally attainable rate of accuracy.

What about AI that is designed to *do* things rather than predict them? I have been talking of AI judges that *decide* legal cases. I have suggested that the ideal

52. I thank an anonymous referee for pressing me to address these further concerns.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

embodiment of this AI would predict what a human would do. But perhaps this suggestion is unreasonably restrictive. It may be tempting to think instead that the ideal embodiment of this AI just decides the cases itself rather than predicting what a human judge would do. This temptation is wrongheaded, however. To understand why, it is important to disentangle reality from metaphor. When we talk of AI deciding to do *xyz*, we are often speaking in metaphor. What we mean is that AI will predict that doing *xyz* will be regarded as the right thing to do by humans. So, when we say that AI decides legal cases, what we must mean is that it spits out an outcome that it predicts humans will regard as correct. Remember, that we are not talking about Artificial General Intelligence (AGI) from science fiction that has a mind of its own.[53] Metaphors aside, all that the (relevant kind of) AI does is predict.

Let me pull all of this together in a relatable example. Consider an AI system that moderates social media posts. Suppose it is tasked to flag hate speech on a social media platform. How does it go about the business of judging which posts to flag? It tries to predict which posts a human moderator would flag as being in violation of a content policy and will go about flagging those. Its 'judging' then consists in no more than predicting what a human moderator would judge in the same circumstances.[54] Its design, therefore, rests on the modest correlative determinism assumption about human flagging behavior, which is to say, its design assumes that whether a human would flag each post in the future will be, at least to an extent, a correlative function of some past events or behaviors. This assumption happens to be, to a reliable extent, true. Experience shows that whether someone would flag a post is to varying degrees a correlative function of various past events. This is why an AI can fulfill this function to varying levels of accuracy. The attained level of accuracy depends on *which* past events or *whose* past behaviors the AI is basing its predictions on.[55] If the AI made its predictions on better predictors, it would make better predictions; and if it made its decisions on the best predictor, which may be a combination of several predicting factors, its accuracy can in principle reach the ideally attainable level of accuracy.

Much current technology research focuses on improving AI performance by fine-tuning training data. Speaking without metaphor, this research focuses on improving the accuracy of AI prediction by discovering better predictors. For instance, Balagopalan et al. have recently discovered a distinction in the strength of two predictors that helps improve the kind of AI we have just been

53. *Cf* Brennan-Marquez & Henderson, *supra* note 2, who imagine the future AI judge to be an AGI. For discussion of this point, see Afrouzi, *supra* note 51 at 30-31.

54. Unless, of course, it is predicting some judgment-independent factual event of the kind I discussed above in section 4.5. See e.g. Jon Kleinberg et al, "Human Decisions and Machine Predictions" (2018) 133:1 QJ Economics 237. Even so, similar assumptions and considerations about (comparative) correlative predictability of past events apply.

55. Röttger et al. show that the presence or absence of annotator subjectivity in the data could make for a better or worse predictor, depending on the predictive task at hand. See Paul Röttger et al, "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks" (29 April 2022), online: *arxiv* arxiv.org/abs/2112.07475.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

considering.[56] The distinction is one between past normative judgments (e.g., judgments about whether something is violative of a rule) versus past descriptive judgments (e.g., judgments about whether something exhibits features that a rule prohibits). Both are reliable predictors of future normative judgments; however, Balagopalan et al. were the first to show that the former is a much better predictor than the latter.

In the context of our social media example, which Balagopalan et al. also discuss, past *normative* judgments are judgments about whether a particular post violates a rule against hate speech. Past *descriptive* judgments are about whether a post exhibits hate speech. Both sets of data could be used to train an AI to make *normative* decisions, that is to say, decide whether a human would regard a new social media post as violative of a rule against hate speech. If trained on descriptive data, the AI will first try to predict whether a human, as a descriptive matter, would regard the post as exhibiting hate speech and then implements the rule with a simple 'if then' logic (hence flagging posts that it predicts humans would regard as exhibiting hate speech). If trained on normative data, the AI will directly try to predict whether a human would, as a normative matter, regard the post as violative of a rule against hate speech (hence flagging posts that it predicts humans would regard as violating a rule that prohibits hate speech).

Both AIs would perform quite well, that is to say, the posts they each flag as violating a rule against hate speech will significantly overlap with those posts that humans would have flagged. But the AI trained on normative data performs much better (the posts *it* flags will have more overlap with those humans would flag). Why? Because, as it turns out, humans tend to be more lenient in their normative judgments than in their descriptive judgments, meaning that they are less likely to say that a post *violates a rule* (against hate speech) than saying that it *exhibits a rule-prohibited feature* (hate speech).[57] As such, the AI trained on descriptive data will incorrectly flag cases where a human would regard the post as exhibiting hate speech though not violative of a rule against it. The AI trained on normative data, meanwhile, will not make such mistakes and will therefore reach a higher level of accuracy in its prediction of human normative judgment.

In short, Balagopalan et al. show that AI can reach a greater accuracy in detecting whether a human would regard a social media post to violate a rule against hate speech if it is trained on data that are labeled as violative/non-violative of the said rule rather than data labeled as exhibiting/non-exhibiting hate speech.[58] Such a finding can be presumably generalized to other contexts, to the extent that humans show a tendency in those other contexts to not find a rule violation with the presence of a rule-prohibited feature below a minimum threshold. In computer science metaphor, this sort of finding will be dubbed 'enhancing

56. See Aparna Balagopalan et al, "Judging facts, judging norms: Training machine learning models to judge humans requires a modified approach to labeling data" (2023) 9:19 Science Advances 1.
57. I suspect this may be due to people associating rule violations with punitive consequences.
58. See Balagopalan et al, *supra* note 56.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

AI's performance by fine-tuning training data.' In abstract terms though, the idea is just that discovering better predictors can improve AI predictions about 'what a human would do'.

Note that whichever type of data the content moderator AI is trained on, its starting assumption is that future human judgment regarding rule violation is a correlative function of *some* past events. In the state of the art, the assumption is that they are a correlative function of past descriptive human judgments. In the improved experimental systems designed by Balagopalan et al., the assumption is that they are a correlative function of past normative human judgments. Both assumptions are to an extent true, though Balagopalan et al. show that the assumption of their experimental systems are to a greater extent true, meaning that future human normative judgments are *more* a correlative function of past human *normative* judgments than they are a correlative function of past human *descriptive* judgments.[59]

The point of all this discussion is this: at the most abstract level, and irrespective of their level of accuracy, *all* AI systems, be they of the present or the future, share the same basic assumption and design. The assumption is what I have called *modest correlative determinacy* with regard to any target domain. The design rests on predicting events in the target domain on the basis of its expected correlation with some past events. The essential differences in various AI systems only lie in *which* past events they rely on more to predict the future events of the target domain.

It should be clear, then, that while advancements in the technology (better data, better labeling, better design, more computational power, or what have you) can increase the accuracy of AI systems' predictions about the universe, including human behavior (and including human *normative* behavior), they could not change the *mode* of AI reasoning. All that such advancements could hope to do is get us closer to what I have called the *ideally attainable rate of accuracy* in AI predictions. To stack the cards in favor of even the AI systems of the future, the argument of this paper assumed *one hundred percent accuracy*, which is in any case, either equal to or greater than the ideally attainable rate of accuracy (depending on whether the target human behavior is one hundred percent correlatively deterministic or not). It should be clear then that no AI, past, present, or future, is immune to the critique of this paper. This is because the critique targets an inherent feature of AI, namely correlative reasoning, which is the very form of reasoning that all AI is enabled by.

5.2 Not Idealizing Human-Run Institutions

There may be good reason, including risk aversion, to hold AI to heightened standards. But this is something we should explicitly deliberate on so that good reasons can be set apart from mere bias. Because absent good reasons, holding AI

59. See Balagopalan et al, *supra* note 56.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

actors (or AI-run institutions) to standards we do not normally hold humans to could risk halting automation when we should not. The danger for irresponsible techno-pessimism arises when we pretend to hold humans to standards that we actually do not, just to show that AI actors fail to meet them. So let us turn to my characterization of human judging. I said that it would be a good thing for judges to make their decisions on the basis of normative reasons rather than purely causal or statistical reasons. Indeed, I said that parties can reject adverse outcomes with a chuckle if the basis for them was solely causal or correlational reasons. But have I relied on a naïve or romanticized picture of human judging? Or is this a standard to which we can hold human judges?

In human-run institutions of judging, many cases are decided primarily by reference to a rule (be it a statute or precedent). Many are decided on procedural rather than substantive measures. So, at least in many cases, judging in human-run institutions is shallow, formalistic, and constrained quasi-mechanically by precedent. But my critique of AI judges is compatible even with such a modest and realistic view of human judging.

To see why, suppose human judgments consisted in nothing more than mechanistic rule-following. Even so, these judgments would be perfectly responsive to normative inquiries in the narrow sense that I claimed judicial decisions must be. This is because rules, even when mechanistically applied, are quintessential examples of normative material that are in the business of justification. And a decision made on the basis of a rule is normative in the sense that it specifies what ought or ought not to be done. Thus, an explanation of the kind: 'This court finds that *xyz* because there is a rule that we should do so' for a decision, though by no means capable of guaranteeing its justification, is sufficient to meet the standard I have set. This is because that kind of explanation is at least in the business of justifying legal judgments, which is all that the standard asks.[60]

But note how the Laplacean scientist could not even do this much: instead of reasoning that he should hold Tom liable for paying Chris because there is a rule that he should so hold, the scientist determines—on the basis of the causal effects of the constellations of physical material and forces—that the human judge would have held Tom liable, had she been presiding. So too, the syntax monster will statistically determine the arrangement of text coming out of the human judge's printer, on the basis of the statistical correlations between all the texts in the world. This means that neither the Martian nor the syntax monster are even up to the task of mechanistic rule-following. Given the argument of the paper, the same goes for all AI systems.

As we saw in section 4.1 above, even explanations that are responsive to normative inquiry may very well fall short of actually justifying their decisions. Mechanistic rule-following is no exception. For one, this sort of explanation can always be contested, *à la* legal positivists, on grounds that the rule was pointless or corrupt. So, I am by no means claiming that mechanistic rules, much less

60. So, too, are less formalistic explanations of the kind 'Tom ought to have followed rule X but he didn't'.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

any other normative material used as the basis of legal judgments, must themselves be justified or justifiable in some further terms. All I am claiming is that they already ring in the right register.

It is extremely important then, to appreciate just how low this paper's argument sets the normative bar: all that the rationale of a judgment needs in order to pass it is having some normative material in its mix, be it a rule, a precedent, or some other legal value. The bar is so low that any real or hypothetical human-run legal system, on any conception of law, could easily meet it. This is indeed what makes this critique incredibly powerful.

Of course, it may be possible to imagine a human-run legal system that also falls prey to the critique. Human judges already make their decisions partly on the basis of non-normative material, many of which are indeed causal or correlative reasons (not all of which are necessarily illicit).[61] And it may be possible for an institution where judges *solely* rely on such bases for their judgments to still count as a *legal* institution on some conceptions of law. The critique only claims that such an institution is also undesirable. The fact that my argument applies such hypothetical legal systems is not a problem for it, because avoiding what I say is undesirable is so easy for human-run decision-making institutions, that any real or hypothetical legal regime could meet it.[62] Yet, it is a feature that no AI-run decision-making institution could possibly avoid. To summarize then:

> 1) The bases for human decisions can easily have some normative elements and often do.
> 2) A decision based on no normative element whatsoever is undesirable.
> 3) The bases for AI decisions will have no normative element whatsoever in their mix and would therefore be undesirable (by 2) in a way that human judgments need not be and often are not (by 1).

In short, in no way does the argument here idealize human judges or existing practice in human-run institutions. Rather, the argument adopts an incredibly modest standard that any judicial system worthy of our attention should meet and then shows why any AI-run institution necessarily fails to meet it.

---

61. For instance, when sentencing, relying on the correlation of a defendant's race with rates of recidivism may be illicit in a jurisdiction, while relying on the correlation of the defendant's crime history with rates of recidivism may not be.

62. Even decisions by a tyrant's whims meet this bar, though might not meet the conceptual bar of 'counting as law' (depending on whom you ask), because someone's whim is *responsive* to normative inquiry, though often failing to justify decisions that have a non-negligible impact on others.

https://doi.org/10.1017/cjlj.2024.17 Published online by Cambridge University Press

Acknowledgments: I am indebted to Stephen Darwall, Robert Post, Jack Balkin, Jonathan Wolff, Christopher Kutz, John Tasioulas, Tom Christiano, Luciano Floridi, John Searle, Barry Stroud, John Zerilli, Felix Steffek, Weyma Lübbe, Christine Chwaszcza, Andy Yu, Matthew Braham, Kiel Brennan-Marquez, Stephen E. Henderson, Markus Rabe, Ali Afrouzi, Julian F. Müller, Jeffrey Kaplan, Eliane Simon, Nikolas Guggenberger, Jane Bambauer, Stefan Arnold, Simone Zurbuchen, William Darwall, Simona Aimar, Christian Maurer, Ezequiel Monti, Jonthahan Gingerich, Rafael Bezerra Nunes, Pinchas Huberman, Artur Pericles Lima Monteiro, Mehtab Khan, Elisabeth Paar, Aparna Balagopalan, Aaron Mendon-Plasek, Agnes Harriet Lindberg, Yan Kai (Tony) Zhou, Samantha Godwin, Marcela Mattiuzzo, Mila Samdub, Gregory Antill, Salwa Hoque, Ça rı Gürkanlı, Vijay Keswani, David Sidi, Daniel Ward, Khomotso Moshikaro, and Daniela Dias. I also presented earlier versions of this paper at the University of Hamburg (2019), University of Regensburg (2019), Cambridge Legal Theory Discussion Group at the University of Cambridge (2020), University of Cologne (2021), Oxford Blavatnik School of Government (2022), Yale Law School (2022), Yale School of Global Affairs (2022), and the University of Arizona Law School (2022). I am indebted to the audiences at all these events. Finally, I am indebted to four anonymous referees for their comments and constructive criticism. All faults remain mine.

Amin Ebrahimi Afrouzi is a Fellow of the Information Society Project at Yale Law School. Email: a.e.afrouzi@yale.edu