ARTICLES

ACCURACY AND THE ROBOT JUDGE

Michael J. Hasday*

Accuracy and efficiency are the two lodestars of civil procedure.[1] But there is an inherent tension between these goals.[2] In the U.S. legal system, that tension has almost always been resolved by making accuracy the paramount goal, so that accuracy is sacrificed only at the

* Assistant Professor, Benjamin L. Crump College of Law at St. Thomas University. B.A. 1997, University of Pennsylvania; J.D. 2001, Northwestern Pritzker School of Law. I would like to thank Michael Abramowicz, Jacob Hamburger, Carol M. Hasday, Jill E. Hasday, Joni Hasday, Robert J. Hasday, Bert Huang, Alan Miller, Siegfried Wiessner, and the participants in the Thirteenth Annual Junior Faculty Federal Courts Workshop and the St. Thomas University College of Law Brown Bag Luncheon Series. I am grateful to the editors of the Journal of Appellate Practice and Process for their thoughtful editing and Erica Lundblad for her excellent research assistance.

1. *See* RICHARD POSNER, ECONOMIC ANALYSIS OF LAW 563 (6th ed. 2003) ("The objective of a procedural system, viewed economically, is to minimize the sum of two types of costs . . . the cost of erroneous judicial decisions [and] . . . the cost of operating the procedural system."); JOHN J. COUND ET AL., CIVIL PROCEDURE 3 (6th ed. 1993) ("There is but one test of a good system of procedure: Does it tend to the just and efficient determination of legal controversies?"); FED. R. CIV. P. 1 (explaining that the Federal Rules of Civil Procedure "should be construed, administered, and employed by the court and the parties to secure the just, speedy, and inexpensive determination of every action and proceeding").

2. Richard A. Posner, *An Economic Approach to Legal Procedure and Judicial Administration*, 2 J. LEGAL STUD. 399, 400 (1973); Patrick E. Longan, *Civil Trial Reform and the Appearance of Fairness*, 79 MARQ. L. REV. 295, 295 (1995) ("Since some procedures that might make the process more accurate would make it less efficient, and some efficient procedures sacrifice accuracy, the goal of reform is to achieve the optimum mix of accuracy and efficiency.").

point of diminishing returns—when a slight gain of accuracy would impose disproportionate cost.[3]

The case for or against robot judges—i.e., a decision maker powered by artificial intelligence (AI)—can be understood within the framework of the accuracy-efficiency tradeoff. If robot judges are less accurate than human judges, or if there is uncertainty about whether robot judges are more or less accurate than human judges, then the question is whether the increased efficiency of robot judges outweighs the lower (or possibly lower) accuracy.

However, the accuracy-efficiency tradeoff has been viewed differently in the robot judge context.[4] Instead of asking how to maximize the accuracy of the legal system to the limit of practicality, analysis of robot judges has focused on the question of whether robot judges are "accurate enough" so that paying for the more expensive human form of judging is no longer worth it.[5]

Indeed, reaching consensus on this question will be difficult if using robot judges to adjudicate disputes could

---

3. *See* Jonathan T. Molot, *Litigation Finance: A Market Solution to a Procedural Problem*, 99 GEO. L.J. 65, 67 (2010) ("A principal goal of civil procedure—indeed, *the* principal goal—is the accurate application of law to fact."); *see also* Paul Stancil, *Substantive Equality and Procedural Justice*, 102 IOWA L. REV. 1633, 1636 (2017); Lawrence B. Solum, *Procedural Justice*, 78 S. CAL. L. REV. 181, 185–86 (2004). There is some debate in the literature whether accuracy can be sacrificed for reasons other than efficiency—such as for the participation rights of the parties. *See* Lawrence Solum, *Legal Theory Lexicon: Procedural Justice*, LEGAL THEORY BLOG (June 4, 2023, 9:00 AM), https://lsolum.typepad.com/legaltheory/2023/06/legal-theory-lexicon-procedural-justice.html [https://perma.cc/JYY9-67MS] (discussing the participation model of procedural justice that "reflects the view that participation matters for reasons other than cost and accuracy"). However, as Solum notes, "[i]n recent years, the question whether participation has value that is independent of outcomes has been enormously controversial." *Id.*

4. *See generally* Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242 (2019) (discussing how artificial intelligence can help and eventually replace human judges).

5. *See id.* at 255–56; *see also* Christopher Michael Malikschmitt, *The Real Future of AI in Law: AI Judges*, AM. BAR ASS'N (Oct. 18, 2023), https://www.americanbar.org/groups/law_practice/resources/law-technology-today/2023/the-real-future-of-ai-in-law-ai-judges/ [https://perma.cc/P9CD-P9XJ] (arguing that the benefit of robot judges is to decide the "many cases that would benefit from a quicker, cheaper, and more mechanical application of the law" and do not require the "haute couture" treatment provided by human judges).

lead to materially less accurate results, which much of the public may be unable to stomach.[6] However, if it can be demonstrated that robot judges are *more* accurate than human judges—and not less—then the case for robot judges has the potential to transcend the accuracy-efficiency tradeoff debate.[7]

But *how* can it be shown that robot judges are more accurate than human judges? Some commentators believe that legal decision making is too indeterminate for this to be demonstrated.[8] Even commentators who believe that this might be possible do not indicate precisely how this might be done.[9] This article provides possible answers to this question, which to my knowledge has not been directly addressed in the burgeoning robot judge literature.[10]

---

6. *See* Benjamin Minhao Chen et al., *Having Your Day in Robot Court*, 36 HARV. J.L. & TECH. 127, 127 (2023).

7. *See id.* at 131.

8. *See* Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1152 (2019); Andrew C. Michaels, *Artificial Intelligence, Legal Change, and Separation of Powers*, 88 U. CIN. L. REV. 1083, 1084 (2019); *see also* JOHN G. ROBERTS, 2023 YEAR-END REPORT ON THE FEDERAL JUDICIARY 6 (2023), https://www.supremecourt.gov/publicinfo/year-end/2023year-endreport.pdf [https://perma.cc/32A4-XTNF] (suggesting that AI cannot replace human judges because "legal determinations often involve gray areas that still require application of human judgment").

9. *See generally* Chen et al., *supra* note 6, at 131 (discussing the possibility of robot judges being more accurate than human judges but not identifying how that might be demonstrated); Adam Unikowsky, *In AI We Trust: AI Is Already Able to Decide Cases Correctly*, ADAM'S LEGAL NEWSL. (June 8, 2024), https://adamunikowsky.substack.com/p/in-ai-we-trust [https://perma.cc/6YXL-LFB5] (suggesting that AI—in its current state of technology—might be more accurate than human judges in deciding cases but leaving to law professors the task of "creat[ing] [the] benchmark that would check whether an AI adjudicator is as accurate as the median human judge, at least in ordinary cases").

10. *See generally* Chen et al., *supra* note 6. Other articles in this literature argue that robot judges might be better than human judges based on an objective other than accuracy. *See* Volokh, *supra* note 8, at 1138–39 (suggesting that robot judges might prove to be more persuasive than human judges based on the metric of expert opinion); Jack Kieffaber, *Predictability, AI, and Judicial Futurism: Why Robots Will Run the Law and Textualists will Like It*, 48 HARV. J.L. & PUB. POL'Y (forthcoming) (manuscript at 8), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4966334 (arguing that robot judges will be better textualists than human judges based on the metric of predictability).

I contend that there are three broad paths through which robot judges can potentially demonstrate superior accuracy over their human counterparts: the consensus path, the comparison path, and the process path.

Part I examines the consensus path, where it is simply accepted by all relevant stakeholders that the robot judge is more accurate than the human judge.[11] While this type of robot judge does not exist in traditional legal settings, it does exist in the professional baseball setting in the form of the robot umpire, which is currently calling balls and strikes at the highest tier of minor league baseball.[12] Beginning the discussion of robot judges in this context allows us to examine the conditions needed for such a consensus to form, provides insight as to whether these conditions could possibly exist in traditional legal settings, and provides a real-world test case about whether a reasonable objection not predicated on accuracy concerns can be made to robot judges.

Part II turns to the comparison path. Here, there is no consensus among all relevant stakeholders, but an argument can be made that the robot judge is more accurate than the human judge based on a comparison using a specified metric.[13] For example, some commentators have contended that robot judges are more accurate than human judges in the bail context based on quantitative metrics for the real-world consequences of the decisions.[14] After describing the

11. *See infra* Part I.

12. *See infra* Part I.

13. *See infra* Part II.

14. *See, e.g.*, Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 133 Q.J. ECON. 237, 241 (2018) ("The algorithmic rule, at the same jailing rate as the judges, could reduce crime by no less than 14.4% and up to 24.7%; or without any increase in crime, the algorithmic rule could reduce jail rates by no less than 18.5% and up to 41.9%. These results are not unique to New York City; we obtain qualitatively similar findings in a national dataset as well."); Jongbin Jung et al., *Creating Simple Rules for Complex Decisions*, HARV. BUS. REV. (Apr. 19, 2017), https://hbr.org/2017/04/creating-simple-rules-for-complex-decisions [https://perma.cc/8BM5-XNRY] ("Despite its simplicity, this rule significantly outperforms expert human decision makers. We analyzed over 100,000 judicial pretrial release decisions in one of the largest cities in the country. Following

three-step framework set forth by proponents of this comparison method, I demonstrate how difficult it is to utilize this comparison method in any particular area of law, including bail hearings. I then introduce a new comparison method that potentially has broader applicability to legal decision making. Under this approach, robot judges can demonstrate greater accuracy than human judges by being better at "matching" the decisions made by human judges deciding appeals from the decisions of lower-level judges.[15]

Part III focuses on the process path. It explores whether an argument can be made for the superior accuracy of the robot judge compared to the human judge based on the process that the robot judge utilizes to reach its decision.[16] I assert that there is a reasonable basis for such an argument, premised on the Condorcet Jury Theorem, when the robot judge replicates the decision of the majority of the qualified human judges.[17] I then consider whether a process-based argument can be made that does not strictly adhere to the Condorcet Jury Theorem. I contend that such an argument can be made if a consensus forms that the process that the robot judge follows is more likely than a human judge to produce accurate decisions. Perhaps venturing into the realm of science fiction (at least for the moment), I then propose a candidate for such a process in the multi-judge context typically found in appellate decision making. In my proposal, a society of robot appellate judges is developed and programmed to deliberate with each other under conditions designed to produce accurate decisions more often than human judges.[18]

---

our rule would allow judges in this jurisdiction to detain half as many defendants without appreciably increasing the number who fail to appear at court."); Sam Corbett-Davies et al., *Even Imperfect Algorithms Can Improve the Criminal Justice System*, N.Y. TIMES (Dec. 20, 2017), https://www.nytimes.com/2017/12/20/upshot/algorithms-bail-criminal-justice-system.html.

15. *See infra* Part II.
16. *See infra* Part III.
17. *See infra* Part III.A.
18. *See infra* Part III.B.

Part IV concludes by briefly considering the effect that robot judges might have on the accuracy of human judges.[19]

I. ACCURACY BY CONSENSUS: THE ROBOT UMPIRE

While there is no robot judge in the traditional legal setting that is widely accepted as more accurate than a human judge, the robot umpire has this distinction in the professional baseball setting.[20] A robot umpire is currently in use in Triple AAA baseball (the highest tier of the professional minor leagues)[21] and might be employed by Major League Baseball (MLB) as soon as the 2026 season.[22] This robot umpire calls balls and strikes through an advanced radar system that tracks pitches through a computer-generated strike zone, adjusted to the height of each batter.[23]

---

19. *See infra* Part IV.

20. I use the term "judge" when referring to the traditional legal setting and "umpire" when referring to the sports setting.

21. Ashley Soriano, *MLB Debuts Robot Umpires for Some Triple AAA Games as Emergence for the Majors Looms*, FOX NEWS (June 20, 2022, 8:45 PM), https://www.foxnews.com/sports/mlb-debuts-robot-umpires [https://perma.cc/Y7SY-5NEV]. MLB is testing in the minor leagues two different ways of using robot umpires. Evan Drellich, *Rob Manfred Goes In-Depth on MLB's Pursuit of an Automated Strike Zone*, THE ATHLETIC (Mar. 29, 2023), https://www.nytimes.com/athletic/4362682/2023/03/29/rob-manfred-automated-strike-zone/ [https://perma.cc/TN87-859B]. In one version, a robot umpire calls every pitch. *Id.* In the other version, a human umpire calls the pitches, but teams can challenge a limited number of pitches for review by a robot umpire. *Id.*

22. *See* Joe Lemire, *MLB Could Test Automated Strike Zone in Spring Training 2025, Manfred Says*, SPORTS BUS. J. (July 16, 2024), https://www.sportsbusinessjournal.com/Articles/2024/07/16/mlb-abs-spring-training-2025/ [https://perma.cc/WN28-9S4H]. While MLB Commissioner Rob Manfred has publicly stated that robot umpires will likely not be used in the Major Leagues for the 2025 season for various reasons, *see infra* notes 71 and 88, MLB is aiming to introduce robot umpires as soon as 2026 with a challenge system and may test the system in spring training games in 2025. *See id.* See *infra* Part I.C.4. for more on the mechanics of the challenge system.

23. Katherine Acquavella, *Robot Umpires*: *How It Works and Its Effect on Players and Managers in the Atlantic League, Plus What's to Come*, CBS SPORTS (Aug. 27, 2019, 5:31 PM), https://www.cbssports.com/mlb/news/robot-umpires-how-it-works-and-its-effect-on-players-and-managers-in-the-atlantic-league-plus-whats-to-come/ [https://perma.cc/E5CK-X4QH]; Jayson Stark, *MLB Just Tweaked Triple A's Electronic Strike Zone: What You Need to Know and Why It*

There is no dispute that the robot umpire is more accurate than the human umpire. While human umpires make mistakes in calling strikes and balls about 8% of the time,[24] robot umpires are much more precise.[25] MLB Commissioner Rob Manfred has flatly stated that robot umpires are more accurate than humans.[26] Indeed, even critics of robot umpires concede they are more accurate.[27] One telling indication of the widespread acceptance of the robot umpire's accuracy is that it is used as the benchmark to evaluate the accuracy of its human counterparts.[28]

It is useful to begin the discussion of robot judges by focusing on robot umpires. First, that starting point establishes the existence of a type of robot judge that is more accurate than its human counterpart. Second, it

---

Matters, THE ATHLETIC (Sept. 7, 2023), https://theathletic.com/4840104/2023/09/07/triple-a-electronic-strike-zone-changes-mlb-explainer/.

24. Davy Andrews, *We May Never Find Out How Good Umpires Can Be*, FANGRAPHS (Feb. 17, 2023), https://blogs.fangraphs.com/we-may-never-find-out-how-good-umpires-can-be/ [https://perma.cc/G8YZ-E9XZ].

25. *See* Zach Helfand, *Invasion of the Robot Umpires*, NEW YORKER (Aug. 23, 2021), https://www.newyorker.com/magazine/2021/08/30/invasion-of-the-robot-umpires [https://perma.cc/6E6J-5L6Z] (noting that "M.L.B. has already concluded that the device is near-perfect, precise to within fractions of an inch").

26. Angelica Stabile, *MLB Commissioner: Robot Umpires Are "More Accurate" than Humans*, FOXBUSINESS (Jan. 22, 2020, 1:27 PM), https://www.foxbusiness.com/sports/mlb-robot-umpires-calls-sports-baseball [https://perma.cc/6U8Z-K4XG].

27. *See* Adam Kilgore, *Analysis: Robot Umpires Would Make Baseball Worse*, PORTLAND PRESS HERALD (Oct. 20, 2021), https://www.pressherald.com/2021/10/20/analysis-robot-umpires-would-make-baseball-worse/ [https://perma.cc/5DJV-YJPH] (arguing that "[b]aseball is better off living with missed calls, as it has for 150 years, than altering the sport in ways it may not even see coming"); John Hirschauer, *Against Robot Umpires*, NAT'L REV. (July 11, 2019, 5:38 PM), https://www.nationalreview.com/2019/07/robot-umpires-baseball-make-game-less-mystical/ [https://perma.cc/G8N4-2JBU] (arguing against robot umpires but conceding that they "are more 'accurate' . . . than their human counterparts"); Rick Morrissey, *Robot Umpires? Let's Leave Baseball to Real, Live Human Beings*, CHI. SUN-TIMES (Jan. 21, 2022, 1:02 PM), https://chicago.suntimes.com/cubs/2022/1/21/22895308/robot-umps-lets-leave-baseball-to-real-live-human-beings-instant-replays-artificial-intelligence-mlb (arguing against robot umpires but conceding they offer "absolute accuracy").

28. Joseph Stromberg, *Robot Umpires Should Be the Future of Baseball*, VOX (July 30, 2015, 6:30 AM), https://www.vox.com/2015/7/30/9068611/baseball-robot-umpire [https://perma.cc/2MVJ-QRFT] ("The MLB implicitly endorses [the robot umpire's] accuracy by using it to grade its umpires after each game.").

allows for an examination of why this form of robot judge is universally considered to be more accurate than its human counterpart, which may have applicability to measuring the accuracy of robot judges in general. Third, this starting point provides a real-world test of whether a reasonable objection can be made to robot judges where it is conceded that the robot judge is more accurate than the human judge.

Of course, a threshold question is whether the "judging" of umpires (at least in the baseball context) is sufficiently analogous to the judging of humans in the traditional legal context for the discussion of robot umpires to be relevant in an article about robot judges. Accordingly, I begin by making the case that the difference between the umpire and judge is one of degree and not of kind.

A. Chief Justice Roberts's Comparison Between Umpires and Judges, His Critics, and Why Both Are Operating Under a Mistaken Premise

Baseball umpires have long been compared to judges, most famously by U.S. Supreme Court Chief Justice John Roberts,[29] who stated in the opening remarks of his confirmation hearing:

> Judges are like umpires. Umpires don't make the rules, they apply them. The role of an umpire and a judge is critical. They make sure everybody plays by the rules. But it is a limited role.
>
> . . .
>
> [I]t's my job to call balls and strikes, and not to pitch or bat.[30]

---

29. *See* Bruce Weber, *Umpires v. Judges*, N.Y. TIMES (July 11, 2009), https://www.nytimes.com/2009/07/12/weekinreview/12weber.html (noting that the comparison between judges and umpires "dots the literature of the 20th century, legal and otherwise" and remarking that Chief Justice Roberts gave baseball umpires "the central metaphorical role in American jurisprudence").

30. *Confirmation Hearing on the Nomination of John G. Roberts, Jr. to be Chief Justice of the United States Before the Comm. on the Judiciary,* 109th Congress 254 (2005) (statement of John G. Roberts, Jr.).

Chief Justice Roberts's comparison has generated criticism on the (perceived) basis that the rules in baseball are broadly agreed upon (e.g., what constitutes a "strike" and "ball" is clear-cut), while the rules in traditional legal settings are a matter of dispute.[31] However, both Chief Justice Roberts and his critics are operating under a mistaken premise. The rules in baseball—and in particular, the scope of the strike zone—have never been free from controversy.[32]

MLB currently defines the strike zone as follows:

> The official strike zone is the area over home plate from the midpoint between a batter's shoulders and the top of the uniform pants—when the batter is in his stance and prepared to swing at a pitched ball—and a point just below the kneecap. In order to get a strike call, part of the ball must cross over part of home plate while in the aforementioned area.[33]

However, the "average" MLB umpire departs from a literal interpretation of the rule in significant ways.[34] For example, although the strike zone is three-dimensional under the rules, pitches that are mostly outside the strike zone during the pitch flight over home plate but "clip the bottom of the three-dimensional zone's front or the top of the back" are usually called balls by human umpires.[35] Although these pitches are arguably "strikes" under the written rule, a "ball" decision for these pitches better conforms to the reasonable expectations of players and fans.[36]

---

31. *See, e.g.*, Theodore A. McKee, *Judges as Umpires*, 35 HOFSTRA L. REV. 1709, 1710 (2007) (arguing that the metaphor obscures the fact that "judges may not be able to systematically decide cases based upon objective application of a set of rules because judges may not agree on what the rules are").

32. *See* Ronald Blum, *What Is a Strike in Baseball? Robots, Rule Book and Umpires View It Differently*, ASSOCIATED PRESS (July 9, 2023, 9:46 PM), https://apnews.com/article/mlb-robot-umpires-strike-zone-40ec7285ae4d1ccaf2621adcb8d72b02 [https://perma.cc/5C8G-S847].

33. *Strike Zone*, MLB, https://www.mlb.com/glossary/rules/strike-zone [https://perma.cc/WBW3-QTYM].

34. *See* Blum, *supra* note 32.

35. *Id.*

36. *See id.*

MLB has struggled with the question of how to program robot umpires when the literal rule and the reasonable expectations of the participants conflict.[37] Indeed, MLB has recently changed the electronic strike zone that is being employed in Triple AAA baseball from three-dimensional to two-dimensional, in contravention of the written rule, to better conform to the reasonable expectations of players and fans.[38]

This same debate—whether to interpret laws and contracts strictly or based on the reasonable expectations of the parties and public—has also been central in American law. Two hypotheticals are instructive. The first is the "No Vehicles in the Park" hypothetical, which has been deemed "the most famous hypothetical in the common law world." [39] In this hypothetical, set forth in 1958, the legal philosopher H.L.A. Hart posed this question:

> A legal rule forbids you to take a vehicle into the public park. Plainly this forbids an automobile, but what about bicycles, roller skates, toy automobiles? What about airplanes? Are these, as we say, to be called "vehicles" for the purpose of the rule or not?[40]

In the decades since Hart first posed this hypothetical, many other hypothetical vehicles have been offered, including an ambulance, a statue of a World War II military truck, a baby stroller, and a wheelchair.[41] This hypothetical has become a staple for courses that introduce students to American law because it crystalizes the issue of whether a rule should be interpreted strictly by the literal meaning of its text or

---

37. *See id.*

38. *See id.* A two-dimensional strike zone is also consistent with the graphic that television producers insert over home plate, which further cements what fans (mis)perceive as the true strike zone. *Id.*

39. Frederick Schauer, *A Critical Guide to Vehicles in the Park*, 83 N.Y.U. L. REV. 1109, 1109 (2008).

40. H.L.A. Hart, *Positivism and the Separation of Law and Morals*, 71 HARV. L. REV. 593, 607 (1958).

41. Pierre Schlag, *No Vehicles in the Park*, 23 SEATTLE U. L. REV. 381 (1999).

should it be interpreted by what a reasonable person would expect the text to mean in context.[42]

The second hypothetical is of more recent vintage and illustrates that this issue remains central to American law. In a 2023 U.S. Supreme Court case,[43] Justice Barrett presented the following scenario:

> [A] parent . . . hires a babysitter to watch her young children over the weekend. As she walks out the door, the parent hands the babysitter her credit card and says: "Make sure the kids have fun." Emboldened, the babysitter takes the kids on a road trip to an amusement park, where they spend two days on rollercoasters and one night in a hotel.[44]

Justice Barrett offered this hypothetical to refute suggestions made by the dissent that she was not being a "good" textualist by departing from what the literal words of the statute at issue appeared to permit.[45] Justice Barrett argued that even a textualist, interpreting the parent's instruction in context with a dose of common sense, would find that the babysitter did not have authorization for her actions.[46] Justice Barrett contended that the trip was not consistent with a reasonable understanding of the parent's instruction and therefore unauthorized, even if it complied with the parent's open-ended instruction in a "literal sense."[47] However, a recent survey of people who are not lawyers suggests that Justice Barrett's view of what constitutes

42. *See* Schauer, *supra* note 39, at 1130–31 ("[T]he example of the no-vehicles-in-the-park rule . . . suggests a real debate about the role of the judge . . . [whether] the good judge is one who sets aside the plain language of the most directly applicable legal rule in the service of purpose, or of reasonableness, or of making law the best it can be, or of integrity, or simply of doing the right thing.").

43. Biden v. Nebraska, 600 U.S. 477 (2023).

44. *Id.* at 513 (Barrett, J., concurring).

45. *Id.* at 534, 539–40 (Kagan, J., dissenting).

46. *Id.* at 513–14 (Barrett, J., concurring).

47. *See id.* (Barrett, J., concurring).

a "reasonable understanding" of the parent's instruction might not be accurate.[48]

B. The Mathematical Formula That Explains Why "Consensus" Accuracy Has Been Achieved for Robot Umpires but Not for Robot Judges

Accordingly, while some may question the relevance of robot umpires to robot judges, the "judging" of robot umpires (at least in the baseball context) has striking similarities (pun intended) to human judging in the traditional legal context. Indeed, we can separate every pitch in the baseball context—and every case in the traditional legal setting—into the same four categories:

Category 1: Determinate

There are pitches where all umpires (both human and robot) who conducted a full review of all the evidence (including video evidence) would agree on whether the pitch was a ball or a strike. For example, all umpires (both human and robot) would agree that a belt high pitch straight down the middle of the plate was a strike.

Similarly, there are cases where all human judges who conducted a full review of all evidence would agree on the decision. For example, these might include an ordinary truck or car driving in the park under no extenuating circumstances in the "No Vehicles in the Park" hypothetical or a babysitter who takes the children to movies and ice cream, just as the babysitter has done many times before without parental complaint, in the babysitter hypothetical.[49]

---

48. *See* Kevin Tobia et al., *Major Questions, Common Sense?*, 97 S. CAL. L. REV. 1153, 1197–1201 (2024) (finding that only 8% of people who are not lawyers thought that the babysitter violated the rule in this hypothetical).

49. Some so-called legal realists contend that every legal case is indeterminate and would not agree with the assertion that there are "determinate" cases. *See* Schauer, *supra* note 39, at 1109. However, this position is difficult to defend in light of clear-cut cases such as those described above. *See id.* (noting that "H.L.A. Hart's example of a rule prohibiting vehicles from a public park was intended primarily as a response to the claims of the legal

Category 2: Indeterminate Based on Law

There are pitches where even after a full review of all evidence (including video evidence) disagreement would persist among the umpires (both human and robot) as to whether the pitch was a ball or strike because of different views about the size of the strike zone. In other words, these pitches are indeterminate because the governing rule for whether the pitch is a ball or a strike is unclear.

Similarly, there are cases where judges who conducted a full review of all evidence would disagree about the appropriate resolution because of differences in judicial philosophy. For example, a strict textualist judge might issue a citation to a child driving a toy car in the park but not find that the babysitter who went on a road trip with the children had violated the parent's instruction. A judge less concerned with the literal words and more open to considering the context might make the opposite decisions.

Category 3: Indeterminate Based on Facts

There are pitches where umpires (both human and robot) who conducted a full review of all evidence (including video evidence) would disagree about whether the pitch was a ball or strike because the pitch was within the fraction of an inch margin of error in the current robot umpire technology. In other words, these pitches are indeterminate because the fact of the location of the pitch is unclear.

---

realists about the indeterminacy of legal rules . . . [who] overestimated law's epiphenomenal indeterminacy and vastly underestimated its everyday determinacy"); *see also* Lawrence B. Solum, *On the Indeterminacy Crisis: Critiquing Critical Dogma*, 54 U. CHI. L. REV. 462, 462, 471–72 (1987) (discussing scholarship that contends that every legal case is indeterminate but arguing that this is demonstrably false because there are easy cases); Brian T. Fitzpatrick, *Many Minds, Many MDL Judges*, 84 L. & CONTEMP. PROBS. 107, 112 (2021) ("[M]ost people seem to agree there are at least *some* legal questions that have determinate answers.").

Similarly, there are cases where differences in factual interpretation produce disagreements among judges who have conducted a full review of all evidence. For example, in the "No Vehicles in the Park" hypothetical, there might be a factual dispute about whether the vehicle actually entered the park. In the babysitter hypothetical, there might be a factual dispute about whether the parent had previously instructed the babysitter to ask the parent for explicit permission if the babysitter was going to spend more than $100.

Category 4: Indeterminate Based on Facts and Law

There are pitches where disagreement persists among umpires (both human and robot) who have conducted a full review of all evidence (including video evidence) because of different views about both the size of the strike zone and the location of the pitch.

Similarly, there are cases where judges disagree after conducting a full review of all evidence because of differences in both judicial philosophy and factual interpretation. For example, in the "No Vehicles in the Park" hypothetical, there might also be a dispute as to whether the toy car was motorized. However, this factual dispute would not be relevant to some judges, whose decision whether to issue a citation to the child would not be affected by whether the toy was motorized.

By separating each pitch or case into these four categories, we can express a circumstance in which "consensus" accuracy exists for the robot umpire or judge with the following formula:

- Let A = the average number of Category 1 pitches or cases per 100 pitches or cases.
- Let B = the average number of Category 2, 3, and 4 pitches or cases per 100 pitches or cases.
- Let C = the percentage of Category 1 pitches or cases that human umpires or judges incorrectly call.

- Let D = the percentage of Category 1 pitches or cases that robot umpires or judges incorrectly call .
- If A x (C – D) > B, there is a consensus that the robot judge is more accurate.

In short, if the average number of mistakes made by human umpires or judges about determinate pitches or cases is greater than the average number of mistakes made by robot umpires or judges about determinate pitches or cases by a number that is more than the average total number of indeterminate pitches or cases, in each case per 100 pitches or cases, then a consensus can form that the robot umpire or judge is more accurate than the human umpire or judge.

For example, assume that the average number of Category 1 pitches or cases per 100 pitches or cases is 95, the average number of Category 2, 3, and 4 pitches or cases per 100 pitches or cases is 5, the percentage of Category 1 pitches or cases that human umpires or judges incorrectly call is 8%, and the percentage of Category 1 pitches or cases that robot umpires or judges incorrectly call is 0%. Since 7.6 (8% of 95) is greater than 5, a consensus can form that the robot umpire or judge is more accurate than the human umpire or judge.

The above formula indicates that the necessary elements for achieving consensus accuracy are: first, a large number of determinate pitches or cases relative to the number of indeterminate pitches or cases; and second, the human umpires or judges making more mistakes than the robot umpires or judges about the determinate cases, assuming the worst-case scenario that a robot umpire or judge is always less accurate than a human umpire or judge with respect to indeterminate pitches or cases.[50] As discussed below, these elements are present in the baseball context. But they are rarely, if ever, found in a traditional legal setting, which is why

50. The assumption that a robot umpire or judge is always less accurate than a human umpire or judge with respect to indeterminate pitches or cases is undoubtedly an overstatement. *See infra* notes 54 and 70 and accompanying text.

consensus accuracy has been achieved for robot umpires but not for robot judges.

With respect to the first element, in the baseball setting, there is a large number of determinate pitches relative to the number of indeterminate pitches because every pitch needs to be called and the vast majority of pitches are not borderline. In contrast, in most traditional legal settings, only cases brought by lawyers will be decided by a judge, and these selected cases are much more likely to be indeterminate than a random case.[51] However, there may be categories of cases, such as asylum cases, foreclosure actions, and patent applications, which are closer to the every-pitch-must-be-called situation in baseball because lawyers do not pick and choose the cases.

With respect to the second element, human umpires in the baseball setting are making a significant number of mistakes about determinate pitches because the speed of professional baseball pitches may be too fast for the human eye to process.[52] While this is a variable that is unique to the baseball setting, perhaps it is not so different from human judges dealing with caseload pressures under which they are forced to make decisions

---

51. This is known as "selection effects" in the literature. *See, e.g.*, Michael Heise & Martin T. Wells, *Revisiting Eisenberg and Plaintiff Success: State Court Civil Trial and Appellate Outcomes*, 13 J. EMPIRICAL LEGAL STUD. 516, 516 (2016) ("Selection effect theory implies that one cannot draw reliable inferences about the larger legal system from studies of tried cases.").

52. *See* Kevin S. Flannagan et al., *The Psychophysics of Home Plate Umpire Calls*, SCI. REPS., Feb. 1, 2024, at 1, 2 (discussing the "demanding perceptual classification task" placed on human umpires given that "the average speed of a MLB pitcher's fastball is 91 to 94 MPH, a baseball takes only 450 milliseconds (ms) to reach the home plate, and is only above the home plate for about 10 ms"). Indeed, better eyesight is likely the reason that younger umpires are significantly more accurate than their more experienced peers. *See* Mark T. Williams, *MLB Umpires Missed 34,294 Ball-Strike Calls in 2018. Bring on Robo-Umps?* BU TODAY (Apr. 8, 2019), https://www.bu.edu/articles/2019/mlb-umpires-strike-zone-accuracy/ [https://perma.cc/K475-AHVU]. Perhaps this is not so different than judges. *See* Ryan C. Black et al., *The Effects of Lifetime Tenure and Aging in the United States Federal Judiciary* 1 (Sept. 19, 2023), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4555766 (finding that "aged judges require more time than their younger colleagues to draft opinions and . . . that older judges are more likely to rely on cognitive shortcuts").

faster than they can do accurately.[53] In contrast, the theoretical promise of robot umpires and judges is that they can be programmed to get any determinate pitch or case right.

However, even if these elements are present to some extent in the traditional legal setting, there are other factors that make it unlikely that a consensus can ever form that robot judges are more accurate than human judges in the traditional legal setting. First, the philosophical differences between umpires on the scope of the strike zone are almost certainly less than the philosophical differences among judges in traditional legal settings, so there will be many more cases that are indeterminate based on law. This is likely due to both the relative clarity of the strike zone rule and the fact that calling balls and strikes is not a politically charged task. Second, as every pitch is recorded with sophisticated video technology, there will be fewer factual disputes in the baseball setting than in traditional legal settings, where most of the "facts" are unrecorded at the time and are filtered through human memory and dependent on human veracity, so there will be many more cases that are indeterminate based on facts.

It is instructive to consider the traditional legal setting of contract interpretation based on ordinary meaning. Judge Kevin Newson of the U.S. Court of Appeals for the Eleventh Circuit, in the context of suggesting that AI-powered large language models (LLMs) might be useful as a tool for human judges for this type of analysis, highlighted an accuracy advantage

53. For example, in a survey of federal judges by the Federal Courts Study Committee, many federal judges admitted that due to caseload pressures, they did not have time to reflect on their own work, read precedential decisions from their own circuit, and even sometimes read the applicable decisions from the U.S. Supreme Court. *See* Lauren K. Robel, *Caseload and Judging: Judicial Adaptations to Caseload*, 1990 BYU L. REV. 3, 57 (1990). While the accuracy-reducing effects of these "short-cuts" in the federal courts might be mitigated somewhat if these tasks are handled by law clerks, it is easy to imagine situations where judges with less ability to hire top-notch clerks will issue inaccurate decisions due to caseload pressures.

this technology has over human judges who rely on dictionaries and their own intuition:

> LLMs are quite literally "taught" using data that aim to reflect and capture how individuals use language in their everyday lives. Specifically, the models train on a mind-bogglingly enormous amount of raw data taken from the internet—GPT-3.5 Turbo, for example, trained on between 400 and 500 billion words—and at least as I understand LLM design, those data run the gamut from the highest-minded to the lowest, from Hemmingway novels and Ph.D. dissertations to gossip rags and comment threads.[54]

Accordingly, one might argue that LLMs have a technological edge for this type of analysis that is similar to the technological edge that robot umpires have in the professional baseball setting. However, even in this discrete legal setting, it is unlikely that a consensus will form that LLMs are more accurate than human judges for reasons including that it is likely that only a small percentage of these cases that are litigated will be determinative cases.[55]

C. A Real-World Test as to Whether Any Criticism of the Robot Judge "Holds Up" Absent Accuracy Concerns

It is difficult to untangle the major criticisms of robot judges from the underlying concerns about the accuracy of robot judges. For example, one objection to the use of robot judges is the so-called black box problem, which is the difficulty in understanding how the robot judge reached its decision.[56] But this objection does not appear particularly salient if it is accepted that the decision is correct. In other words, if we can have confidence in the

54. Snell v. United Specialty Ins. Co., 102 F.4th 1208, 1226 (11th Cir. 2024) (Newson, J., concurring).

55. Of course, a scenario where the robot judge *might* have similar or greater accuracy than a human judge still could be very attractive to particular parties given the efficiency gains.

56. *See* David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS L. REV. 653, 706 (2017).

result, then the need to understand the particular "code" that led to that result loses its importance.[57]

Another objection to robot judges is that they will merely replicate the past biases of human judges.[58] This objection, however, is essentially an accuracy critique, suggesting that robot judges cannot be more accurate than human judges because they will inevitably rely on past human judicial decisions in reaching their legal conclusions. Therefore, this critique does not apply in situations where it is taken as a given that the robot judge *is* accurate.

A third criticism of robot judges is that the public will not view the process as fair if the decision maker is not a human.[59] However, a recent study found that this public view is predicated on concerns that the robot judge is less accurate than the human judge, and not anything inherent about whether a human or machine is making the decision.[60] Indeed, the experience of robot umpires in professional baseball shows that the stakeholders—the players, managers, teams, and fans—more readily accept the decision of the machine over the human. While some commentators argue that there is a "human element" that is lost with robot umpires,[61] this sentiment appears to be at least partly based on the entertainment value that is created by the mistakes (real or perceived) by the human umpires.[62] These mistakes can sometimes

57. In addition, human judges are in a way a "black box" too and perhaps even less "knowable" than a robot judge. *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633, 634 (2017) ("The implicit (or explicit) biases of human decisionmakers can be difficult to find and root out, but we can peer into the 'brain' of an algorithm.").

58. *See, e.g.*, Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2221 (2019).

59. *See, e.g.*, Re & Solow-Niederman, *supra* note 4, at 264.

60. *See* Chen et al., *supra* note 6, at 131 (concluding that its study "suggests that the human-AI fairness gap is explained by 'hard' factors, like the perceived accuracy and thoroughness of the decision-making process, more so than by distinctively human, 'soft' factors, like the decision-maker's understanding of the litigant's position or a feeling that the litigant had a voice").

61. *See* Morrissey, *supra* note 27 ("Real umps are human. That means they're prone to dye jobs, bad breath and errors in judgment, like the rest of us.").

62. *See id.* ("A strike should be a strike, not a pitch three inches off the outside corner. But when I think about artificial intelligence in baseball, I always come

lead to heated and entertaining arguments.[63] The fact that these types of disputes have been largely eliminated by the use of robot umpires is a strong indication that the robot umpire's decision making is viewed as more legitimate than the decision making of human umpires.[64]

However, the debate over the use of robot umpires in professional baseball raises other concerns that might provide the basis for a reasonable objection to a robot judge in a traditional legal setting that is not predicated on accuracy concerns. I call these the too-accurate critique, the good bias critique, the collateral effects critique, and the front door/single decision maker critique. Each of these critiques will be addressed in turn.

1. The Too-Accurate Critique

One criticism about robot umpires is that they are "too accurate"—that while their calls may be technically correct, they do not always conform to the expectations of the average player or fan.[65] As noted above, the use of robot umpires in professional baseball has resulted in some technical "strikes" that appear to the human eye to be balls.[66]

---

back to entertainment. Baseball has lost its ability to hold an audience, and, as silly and illogical as it might seem, flesh-and-blood umpires do have entertainment value.").

63. *See id.* Morrisey also argues that the use of robot umpires making line calls in professional tennis has eliminated much of the argument "theater" between players and umpires, and notes that if this technology existed in the 1970s and 1980, the public would have been deprived of the infamous "out-of-his-mind outraged at the incompetence of line judges" antics of the professional tennis player John McEnroe. *Id.*

64. Helfand, *supra* note 25 (noting that after the minor leagues implemented the robot umpire "the arguments basically stopped").

65. *See* Kilgore, s*upra* note 27 (noting that "[c]ertain rule book strikes have always been called balls, to the acceptance of all parties involved").

66. *See id*. The too-accurate critique has also been made with the use of instant replay in MLB because base stealers who appeared to the naked eye to have been successful were now being called out because the replay cameras captured that the base stealers' slides carried them off the base for a fraction of a second. *See* Billy Witz, *Questions in Baseball Over Unintended Consequence of Instant Replay*, N.Y. TIMES (Oct. 18, 2015), https://www.nytimes.com/2015/

Similarly, a common criticism of robot judges is that they will engage in a literal interpretation of rules, statutes, and contracts while only human judges can take "context" (however defined) into account. For example, Justice Kagan, in criticizing Chief Justice Roberts's umpire-judge comparison, stated that it wrongly implied that judges are "robotic" in applying "clear cut" rules.[67] In a similar vein, critics of a 2020 U.S. Supreme Court decision[68] contended that the majority opinion was engaging in a too-literal reading of the pertinent statute and the analysis had an "algorithmic feel to it."[69] In other words, the Justices who joined the majority opinion were accused of acting like robots.

However, as the experience of the robot umpire shows, a robot judge can be programmed to depart from the literal interpretation of the rules to account for the reasonable expectations of the participants. Furthermore, David Hoffman and Yonathan Arbel have shown that AI has the capability of interpreting contracts not just literally but with the ability to "ascertain ordinary meaning in context, quantify ambiguity, and fill gaps in parties' agreement."[70]

A more complicated situation may arise when there is a lack of consensus among the stakeholders as to whether a certain pitch should be called a ball or strike. When there is a true divide with respect to a pitch, one could argue that a consistent decision made by a robot umpire may be less accurate than an inconsistent decision made by a human umpire.

10/19/sports/baseball/questions-in-baseball-over-unintended-consequence-of-instant-replay.html.

67. Nathan Koppel, *Are Judges Umpires*? *Eh, Not Exactly, Says Kagan*, WALL ST. J. (June 30, 2010, 11:54 AM), https://www.wsj.com/articles/BL-LB-30831 (quoting Justice Kagan as criticizing the metaphor "because it wrongly implies that high-court judging 'is a kind of robotic enterprise . . . that everything is clear cut'").

68. Bostock v. Clayton Cnty., 590 U.S. 644 (2020).

69. *See* Tara Leigh Grove, *Which Textualism?*, 134 HARV. L. REV. 265, 281 (2020).

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

This situation is not just theoretical. Even after changing the electronic strike from three dimensional to two dimensional, MLB is still debating whether further modifications to the strike zone are necessary, particularly in regard to pitches that hit the corners of the strike zone, which are often called "balls" by human umpires.[71]

Moreover, the split in the decision making on these pitches might be a feature, and not a bug, of the system. Because these pitches are very hard to hit, it might encourage pitchers—even those with the control to hit the corners—an incentive to throw more hittable pitches, which arguably has systematic benefits for the game. At the same time, by sometimes calling these pitches strikes, it might keep the strike zone large enough so that pitchers without pinpoint control can still be effective. [72]

Similarly, in the traditional legal setting, there are many areas of the law where the rules are a matter of contention and an argument can perhaps be made that inconsistent decisions are more accurate in the aggregate than consistent decisions. Moreover, there might be societal advantages with parties not knowing precisely where the line is between legal and illegal conduct.

---

71. *See* Drellich, s*upra* note 21 (quoting the MLB Commissioner as stating that "[t]here's a reason [human umpires] don't call [corner pitches]: because you can't hit that"). This ongoing debate was one of the reasons cited for MLB's decision to table robot umpires in the Major Leagues for the 2025 season. *See* Evan Drellich, *Robot Umpires in MLB Could Have 'Unintended Consequences' to Strike Zone, Manfred Says*, THE ATHLETIC (May 23, 2024), https://www.nytimes.com/athletic/5516189/2024/05/23/automated-strike-zone-consequences-manfred/ (quoting the MLB Commissioner as stating that MLB has not "settled" on what it thinks should constitute the strike zone for purposes of programming robot umpires).

72. In other sports, such as football, it is often claimed that certain infractions occur on nearly every play and that it would ruin the flow of the game if the foul was always called. *See, e.g.*, Joseph Zucker, *Tom Brady Says NFL Referees Could Call Holding Penalty on 'Every Single Play*,' THE BLEACHER REP. (Jan. 31, 2023), https://bleacherreport.com/articles/10063607-tom-brady-says-nfl-referees-could-call-holding-penalty-on-every-single-play [https://perma.cc/424U-Q9QF] (quoting the NFL legend as stating that "there [is] holding by the offensive line and defensive line on every play").

However, this is not a reasonable objection to the robot judge because, to the extent there is a consensus that a judge should make a certain decision only some of the time, the robot judge can be programmed to decide cases/pitches on a percentage basis. For example, in the baseball setting, the robot umpire could be programmed to call pitches that hit the corners strikes 60% of the time and balls 40% of the time if that breakdown is considered optimal.[73]

In short, the too-accurate critique relies on a mistaken premise and cannot constitute a reasonable basis to object to the robot judge.

2. The Good Bias Critique

One of the perceived advantages of robot umpires and judges is the potential for bias-free judgments. For example, human umpires are subject to a plethora of human biases, including the gambler's fallacy,[74] anchoring effects,[75] impact aversion,[76] and even racial

73. In a similar vein, Adam Unikowsky argues that "if we use AI [to decide legal cases], we can add exactly as much unpredictability and ideology as we want via effective prompt engineering rather than having unpredictability thrust upon us by the constraint of individual judges' bandwidth." *See* Unikowsky, *supra* note 9.

74. *See* Keith Law, *Human Fallibility and the Case for Robot Baseball Umpires*, WIRED (May 1, 2020), https://www.wired.com/story/human-fallibility-case-robot-baseball-umpires/ [https://perma.cc/7WWB-Q5MP]. The gambler's fallacy is the mistaken belief that effectively random events will "even out" in a small sample. *Id.* An example is the unfounded belief that if a roulette ball has landed on red three times in a row, it is more likely to land on black the next time because black is now "due." *See id.* Studies have shown that umpires are more likely to call a "strike" if their call on the previous pitch was a "ball" and vice versa. *Id.*

75. *See id.* Somewhat different than the gambler's fallacy, anchoring effects is the cognitive bias of allowing the first piece of information received to disproportionately affect subsequent decisions. *See id.*

76. *See id.* Impact aversion is a bias to preserve the status quo of the game. *See id.* The prime example is that human umpires are less likely to call a "strike" if the batter already has two strikes (causing a strikeout) and less likely to call a "ball" if the batter already has three balls (causing a walk). *See id.*

bias.[77] However, some believe that certain biases of human umpires and judges are a positive.

For example, it is well known that human umpires are more likely to call strikes on borderline pitches if the pitcher has a reputation for accuracy.[78] Some commentators have suggested the star pitcher has "earned" the benefit of the doubt on these pitches.[79] There is (perhaps) an argument that this type of bias improves the game overall as it incentivizes pitchers to be accurate, which increases the entertainment value of the game.

In a similar vein, some have suggested that an advantage of human umpires is that they can enlarge the strike zone when there is a reason to speed up a game, such as when the ultimate outcome of the game is not in doubt because of a lopsided score.[80]

---

77. *See* Hank Snowdon, *Would "Robot Umpires" Reduce Discrimination? Measuring Racial Bias in Major League Baseball Umpires*, CMC SENIOR THESES ii (2021) (finding that "umpires are significantly more likely to make calls that favor players of the same race, and that these effects have not diminished between 2008 and 2020").

78. *See* Nicholas Bakalar, *Ball? Strike? It Depends: Is the Pitcher an All-Star?*, N.Y. TIMES (July 17, 2014), https://www.nytimes.com/2014/07/08/sports/baseball/study-finds-umpires-ball-strike-calls-favor-all-star-pitchers.html ("For each additional appearance in an All-Star Game there was a 4.8% increase in the probability that an actual ball would be called a strike. A player with five All-Star appearances had a 14.9% chance of a true ball being called a strike, which is a 16.7% increase over the chance a journeyman will benefit from the same mistake.").

79. *See* Kilgore, *supra* note 27 (lauding star pitchers with "[t]he ability to manipulate an umpire's human set of eyes . . . [by] finding out what they could get away with and slowly inching pitches further out, making the umpire widen his zone without even realizing it"); Bakalar, *supra* note 78 (quoting a former MLB umpire stating that "[i]f a pitcher is throwing strikes, then it's accepted that the zone is expanded," but "[i]f he's not, he's got to throw a defined strike").

80. *See* Jayson Stark, *Are Robot Umpires Ready for Their MLB Debut? Not So Fast*, THE ATHLETIC (Aug. 25, 2023), https://www.nytimes.com/athletic/4791440/2023/08/25/mlb-robot-umpires-future/ (asking if minor league players, coaches, and managers "want every pitch to be called by non-humans, strictly by the rulebook strike zone, in, say, a 17-2 game? Or when it's the eighth inning and ominous dark clouds are gathering?"). In other sports, some commentators believe that certain infractions should be called less often, or perhaps not at all, during the closing minutes of the game, so the players, and not the referees, "decide the game." *See, e.g.*, Sabreena Merchant, *Controversial Late-Game Foul Against UConn's Aaliyah Edwards Draws Ire: 'I Wasn't Given an Explanation*,*'* THE ATHLETIC (Apr. 5, 2024), https://www.nytimes.com/athlet

In the traditional legal context, attorneys with a good reputation for thoroughness and honesty built over many years may get the benefit of the doubt (not given to other attorneys) by human judges for assertions made in court.[81] One indication that this is a type of "bias" is considered to be the "good kind" by some in the legal profession is that it is openly discussed by both attorneys and judges.[82]

Moreover, human judges may treat pro se plaintiffs differently than plaintiffs represented by attorneys. There is a public debate among judges about whether the procedural treatment of pro se plaintiffs should be more lenient or exactly the same as the procedural treatment of represented plaintiffs.[83] However, there appears to be a behind-the-closed-doors debate among judges as to

---

ic/5395000/2024/04/06/uconn-foul-aaliyah-edwards-iowa-final-four/ (discussing the controversy over a late-game foul call, but noting that "[i]t's a call that wouldn't have drawn much ire in the first quarter").

81. *See, e.g.*, Kevin T. McGuire, *Repeat Players in the Supreme Court: The Role of Experienced Lawyers in Litigation Success*, 57 J. POL. 187, 189 (1995) (arguing that repeat player attorneys get better results at the U.S. Supreme Court because the Justices know they have more reputational incentive to present more accurate information in a given case and therefore have more "credibility" with the Justices).

82. *See, e.g.*, Rebecca L. Palmer, *Attorney-Judge Relationship Is a Significant Factor in Court*, REBECCA L. PALMER LAW GRP. (Aug. 16, 2022), https://rlplawgroup.com/attorney-judge-relationship-is-a-significant-factor-in-divorce-court# [https://perma.cc/L39V-BRW8] ("Having a reputable attorney on your team also influences your case's outcome. A judge will recognize if your attorney is experienced, respectable in the field, and trustworthy. Working with a lawyer who has proven themselves in their field and has established a strong reputation is extremely important."); Cedra Mayfield, *'I Would Never Trust That Lawyer Again': Judges Discuss How Attorneys Damage Their Own Reputations*, LAW.COM (Sept. 7, 2021, 4:22 PM), https://www.law.com/dailyreportonline/2021/09/07/i-would-never-trust-that-lawyer-again-judges-discuss-how-attorneys-damage-their-own-reputations/ [https://perma.cc/U68L-PHEF] (quoting a former judge as stating: "Be careful what you do in front of the judge. You might get an advantage on that case, but it's not worth it, ultimately, for your reputation or your career."); *How to Establish a Positive Courtroom Reputation with Judges*, STATE BAR OF TEX.: TEX. BAR BLOG (June 21, 2016), https://blog.texasbar.com/2016/06/articles/news/positive-courtroom-reputation-with-judges/ [https://perma.cc/2532-62C2] (discussing a state bar panel discussion where judges stressed the importance of attorney reputation).

83. *See generally* Julie M. Bradlow, *Procedural Due Process Rights of Pro Se Civil Litigants*, 55 U. CHI. L. REV. 659 (1988) (summarizing the debate and arguing that pro se litigants should be treated more leniently).

whether pro se plaintiffs should be given less judicial attention than represented plaintiffs, perhaps suggesting that the procedural hurdles for the pro se plaintiff to get their proverbial day in court might be higher than for represented plaintiffs.[84]

As in the baseball context, there are arguments that these types of bias are the "good kind" and an overall net positive for the legal system. For example, providing an advantage to attorneys with a strong reputation may incentivize good behavior among attorneys. Furthermore, it could be argued that relaxing the procedural requirements of civil procedure for pro se litigants makes the litigation fairer by ensuring that the dispute is decided on the merits and not technicalities. Somewhat paradoxically, a judge might dismiss a pro se complaint, even if the complaint technically complies with the governing pleading standard, because of a belief that the pro se plaintiff is unlikely to be able to prove her case without the assistance of counsel, and therefore dismissing the case early will save judicial resources without impacting the outcome of the case.

This objection fails for the same reason as the too-accurate objection: to the extent that any bias is desired, the robot umpire or judge can be programmed to exhibit that bias. However, it is highly unlikely that MLB or a judicial system would explicitly agree to program a robot umpire or judge to exhibit any bias. Furthermore, even if MLB or a judicial system would agree, there is no consensus that these arguably good biases are actually

84. *See, e.g.*, David Lat, *Judge Posner, Uncensored: "I Don't Really Care What People Think,"* ABOVE THE L. (Sept. 14, 2017, 7:41 PM), https://abovethelaw.com/2017/09/judge-posner-uncensored-i-dont-really-care-what-people-think/ [https://perma.cc/Z5BR-ABM4] (quoting the recently retired Seventh Circuit judge as stating: "When pro se litigants appeal, their appeal papers are given to a staff attorney. We have about 20 staff attorneys who are appointed for two years, and a few supervisors. The staff attorneys tend to be good students from good schools, hired right after they graduate. Despite their good credentials, they tend to be hostile to the pro se's. It's not their own feelings; it's that they sense—correctly—that the judges don't really care much about the pro se's, find them nuisances, and are not interested in them. So that percolates down to the staff attorneys, and they have a tendency to go against the pro se appeals even when they have apparent merit.")

"good." Indeed, biases around pro se plaintiffs may point in opposite directions depending on the values and priorities of the judge.

3. The Collateral Effects Critique

With any change, there will be winners and losers, and some skills that have been developed (or not developed) at great cost may no longer be valued. With robot umpires, the most prominent skill that will become obsolete is pitch framing. The precise definition of pitch framing is under some dispute, but a good working definition is this:

> Framing is a method of receiving the pitched ball from the pitcher, using subtle movements of the catcher's wrist and body, made for the purpose of presenting the pitch to the umpire in a manner which increases the likelihood that the pitched ball will be called a strike.[85]

In other words, catchers have developed a skill to essentially trick the umpire into believing a "ball" is a "strike." And the value of this skill should not be underestimated: by some sophisticated statistical accounts, a catcher who is an expert in this "dark art" can "save his team as many as 25 runs more than the average catcher, over the course of the season [which] may be worth as many as an additional 2–3 wins, relative to the average catcher."[86] Some critics of robot umpires cite the loss of this skill as a reason to oppose

85. Sheryl Ring, *Is Pitch-Framing Cheating?*, FAN GRAPHS (July 27, 2018), https://blogs.fangraphs.com/is-pitch-framing-cheating/ [https://perma.cc/22ZH-MUA7]. The official MLB definition is more diplomatic but essentially says the same thing. *See Catcher Framing*, MLB, https://www.mlb.com/glossary/statcast/catcher-framing [https://perma.cc/8TNV-A79Q] ("Catcher framing is the art of a catcher receiving a pitch in a way that makes it more likely for an umpire to call it a strike—whether that's turning a borderline ball into a strike, or not losing a strike to a ball due to poor framing.").

86. Sameer K. Deshpande & Abraham Wyner, *A Hierarchical Bayesian Model of Pitch Framing*, 13.3 J. QUANTITATIVE ANALYSIS SPORTS 95, 95 (2017).

robot umpires.[87] Even MLB Commissioner Manfred has expressed concern about this "unintended consequence" of implementing robot umpires.[88]

Indeed, with robot umpires, the catchers who have developed this skill at great cost will suddenly be much less valuable to their teams and either be compensated at a much lower level or lose their jobs altogether.[89] Similarly, a lawyer's skill at framing an argument to lead the human judge to the "wrong" decision (but "right" for her client) will become obsolete with a robot judge. Accordingly, much like catchers, attorneys who have this skill will become less valuable and can expect lower compensation.[90] Are these collateral effects on catchers and attorneys something we should worry about?

I believe the answer is no. While the ability of catchers to frame pitches will have less value under a regime of robot umpires, it is difficult to argue that baseball should want to encourage catchers to develop the ability to manipulate umpires into making incorrect

---

87. *See* Kilgore, *supra* note 27 (arguing that "[t]he ability to manipulate an umpire's human set of eyes is part of the game's artistry"); Hirschauer, *supra* note 27 ("Let the catchers keep framing pitches, for goodness' sake.").

88. *See* Drellich, *supra* note 71 (quoting the MLB commissioner as stating: "It's the unintended consequences of [robot umpires]. The one that is often pointed to, but not the only one, is the framing catcher. . . . I mean, that alters peoples' [sic] careers. Those are real, legitimate concerns that we need to think all the way through before we jump off that bridge.").

89. *See* Stark, *supra* note 80 ("If the sport goes to full-time ball/strike technology, framing will cease to have value. Period. Which means the league would be telling men who have devoted years to that part of their craft: *Hey, thanks for your time—but now go find something else to be good at.*"); *see also* Taylor Bechtold, *AI Takeover, Part I: Will Some Catchers Be Pushed Out of Baseball When the Robot Umpires Arrive?*, OPTA ANALYST (Aug. 31, 2022), https://theanalyst.com/na/2022/08/will-some-catchers-be-pushed-out-of-baseball-when-the-robot-umpires-arrive/ [https://perma.cc/RV6K-4XBS]. On the other hand, there might be non-monetary benefits to catchers if this practice is retired. *See* Katie Woo, *Catcher's Interference Calls Are Skyrocketing in MLB. It's Putting Players at Risk*, THE ATHLETIC (May 9, 2024), https://www.nytimes.com/athletic/5480350/2024/05/09/cardinals-mlb-willson-contreras-catcher-risk/ (discussing the rise of catcher injuries as their attempts to pitch frame causes them to move closer to plate, which increases the likelihood of their being struck by the bat).

90. *See* Re & Solow-Niederman, *supra* note 4, at 274. In a similar vein, Re and Solow-Niederman write that "[t]he power of a lawyer's rhetoric . . . would count for much less in a legal system where AI adjudicators are capable of ruling on thousands of technically drafted motions for summary judgment." *See id.*

decisions for the underlying purpose of having those catchers receive more compensation. Similarly, while the ability of attorneys to frame arguments will have less value under a regime of robot judges, it is difficult to argue that society should want to encourage attorneys to develop the ability to manipulate judges into making incorrect decisions for the underlying purpose of having those attorneys receive more compensation.

4. The Front Door/Single Decision Maker Critique

I will conclude with what I believe is the strongest critique of robot umpires, which I call the front door/single decision maker critique. While this critique has not been explicitly made, it is implicit in much of the criticism discussed above. The argument is that while robot judges *could* of course be programmed in all sorts of ways, it is not publicly acceptable to explicitly acknowledge, for example, that the strike zone becomes larger in lopsided games even though there might be a private consensus that it should. Accordingly, we need human umpires to get through the "back door" what robot umpires are unable to get through the "front door."

A related argument is that it is too difficult for a single decision maker like MLB to determine the optimal strike zone and we might have more confidence in a decentralized process involving many different human umpires. If MLB makes a mistake in how it programs the robot umpire, the negative impact on the game will be much larger than any mistake made by an individual human umpire.

Accordingly, under this argument, we should want individual human umpires deciding every *indeterminate* pitch even if these decisions are inconsistent or biased. This does not mean that human umpires should be calling pitches "balls" that are clearly "strikes" and vice versa. It does mean, however, that human umpires can enlarge or narrow the strike zone within the "indeterminacy range" to account for a variety of factors. For example, in a lopsided game, it might be appropriate for an individual human umpire to enlarge their strike

zone as long as there is at least a colorable claim that the pitches at the edges of the enlarged strike zone are still strikes.

However, even if you accept the view that we want human umpires calling every indeterminate pitch, which I do not, it is not an argument against robot judges in toto; it is only an argument against robot judges calling indeterminate pitches. Indeed, it appears that the first version of robot umpires that will be used in MLB will be an effort to have the human umpires calling the indeterminate pitches and the robot judges calling (or at least controlling) the determinate pitches. Under the "challenge system" being considered, the human umpire makes the first call on each pitch, but each team has a small number of challenges (two or three) to appeal the human umpire's decision to the robot umpire, which then makes the final call.[91] The pivotal rule under this system is that the team retains its challenge when it is proved correct but loses its challenge when it is proved wrong.[92] This rule therefore incentivizes teams to only challenge calls that are clearly wrong—i.e., the human umpire's mistakes on determinate pitches—because this allows the team to make numerous challenges.[93]

---

91. *See* Jayson Stark, *Triple A-Games to Start Fully Using Automated Ball-Strike Challenge*, THE ATHLETIC (June 18, 2024), https://www.nytimes.com/athletic/5573707/2024/06/18/automated-ball-strike-challenge-system-triple-a/. Assuming the challenge system is implemented in MLB, it is unclear whether it will be adopted permanently or as a "stepping stone" to a system in which the robot umpire makes the decision on every pitch. Dayn Perry, *MLB Will Reportedly Pivot to ABS Challenge System in Triple-A, After Rob Manfred Teased Future of Robo Umps*, CBS SPORTS (June 18, 2024, 1:59 PM), https://www.cbssports.com/mlb/news/mlb-will-reportedly-pivot-to-abs-challenge-system-in-triple-a-after-rob-manfred-teased-future-of-robo-umps/ [https://perma.cc/HA82-3DYV].

92. *See* Stark, *supra* note 91.

93. *See id.* (noting that under the challenge system, robot umpires are "used only to correct the most egregious mistakes"); Jesse Rogers, *When, How Will Robot Umps Arrive in MLB? Latest on ABS Plans*, ESPN (June 18, 2024, 1:50 PM), https://www.espn.com/mlb/story/_/id/40377683/mlb-robot-umpires-automated-balls-strikes-challenge-system-umps-majors [https://perma.cc/48YA-L2F6] (noting that "[s]ince teams retain their number of challenges if they win, the actual number used in [a minor league game] has reached double digits").

In the traditional legal setting, we can imagine something analogous to the challenge system so that robot judges are deciding (or at least controlling) the determinate cases but human judges are deciding the indeterminate cases. For example, as in the challenge system, human judges could initially decide each case with the possibility of an appeal to a robot judge. However, if the robot judge affirms the human judge's decision, then the party that sought the appeal would be responsible for paying the other party's attorney fees or another specified penalty. This would incentivize parties to only appeal decisions where the human judge was clearly wrong, i.e., where the human judge decided a determinate case incorrectly.

In sum, the real-world test case of robot umpires provides no support for the proposition that a reasonable objection can be made to the robot judge that is not predicated on accuracy concerns.

II. ACCURACY BY COMPARISON

In the traditional legal context, there is currently no robot judge that by consensus is considered to be more accurate than a human judge. However, some commentators have argued, based on a comparison using a specified metric, that robot judges are currently more accurate than human judges in certain contexts.[94] The three-part framework I am presenting below has not been explicitly set forth in the literature, but the comparison argument appears to be made in three parts. First, proponents of this argument contend there is a consensus as to the legal objective of a legal proceeding in a particular context.[95] I will refer to this part of the argument as the "Consensus on the Legal Objective." Second, proponents contend that there is a specified metric to measure the ability of human judges or robot

94. *See, e.g.*, Kleinberg et al., *supra* note 14, at 241.

95. *See, e.g.*, *infra* note 100.

judges to achieve the Consensus on the Legal Objective.[96] I will refer to this part of the argument as the "Specified Metric." Third, proponents contend that it can be demonstrated empirically that robot judges are better at achieving the Specified Metric than human judges.[97] I will refer to this part of the argument as the "Empirical Demonstration."

Some commentators use this type of comparison argument to contend that the robot judge is currently more accurate than the human judge in certain areas of criminal law, such as bail hearings.[98] Indeed, this argument has convinced jurisdictions across the country to utilize pre-trial risk assessment algorithms—in essence, robot judges—to effectively decide many of these cases, even though human judges have the right to override the robot judges' decisions.[99]

The argument that the robot judge is more accurate than the human judge in the bail hearing context proceeds as follows:

Step 1: Consensus on the Legal Objective

Proponents contend that there is a consensus that the legal objective of bail hearings is to limit the risk of "flight" and "danger" while minimizing the amount of time arrested persons spend in prison pre-trial.[100]

---

96. *See, e.g.*, *infra* note 101.

97. *See, e.g.*, *infra* note 102.

98. *See* Kleinberg et al., s*upra* note 14, at 241.

99. *See* Ngozi Okidegbe, *Discredited Data*, 107 CORNELL L. REV. 2007, 2009 (2021) ("[J]urisdictions across the country are increasingly adopting pretrial risk assessment algorithms.") (citing PRETRIAL JUST. INST., THE STATE OF PRETRIAL JUSTICE IN AMERICA 3, 13 (2017) ("25% of people living in the United States now reside in a jurisdiction that uses a validated evidence-based pretrial assessment.")); *see also* Victoria Angelova et al., *Algorithmic Recommendations and Human Discretion* 12 (Nat'l Bureau of Econ. Rsch., Working Paper No. 31747, 2023) (finding that judges accept the recommendation of the algorithm in the jurisdiction studied (a large, Mid-Atlantic city) about 80% of the time).

100. *See* Okidegbe, *supra* note 99, at 2009 n.1 ("In terms of pretrial misconduct, the bail system is designed to release defendants except those posing a risk of non-appearance, obstruction of justice, and, in most jurisdictions, danger to public safety."); Note, *Bail Reform and Risk Assessment: The*

Step 2: Specified Metric

Proponents contend that the specified metric to measure the ability of human judges or robot judges to achieve the Consensus on the Legal Objective is, at any given release rate, how often the arrested persons fail to appear at the court hearing ("limit flight") or are re-arrested before trial ("limit danger").[101]

Step 3: Empirical Demonstration

Proponents contend that it can be demonstrated empirically that robot judges are better at achieving the Specified Metric than human judges.[102]

However, many other commentators dispute the claim that robot judges are more accurate than human judges in the bail hearing context, by contesting one or more of the steps above. First, some commentators take issue with Step 1, arguing, for example, that an equally important legal objective of bail hearings is to eradicate racial disparities.[103]

Cautionary Tale of Federal Sentencing, 131 HARV. L. REV. 1125, 1126–27 (2018) ("In most jurisdictions, a person may be detained pretrial only if there is a high risk that the person will not appear in court or that the person will be a danger to the community before trial." (citing CRIMINAL JUSTICE POLICY PROGRAM, HARVARD LAW SCH., MOVING BEYOND MONEY: A PRIMER ON BAIL REFORM 5–6 (2016))); Laura I. Appleman, *Justice in the Shadowlands: Pretrial Detention, Punishment, & the Sixth Amendment*, 69 WASH. & LEE L. REV. 1297, 1330 (2012); *see also* United States v. Salerno, 481 U.S. 739, 755 (1987) ("In our society liberty is the norm, and detention prior to trial or without trial is the carefully limited exception.").

101. *See, e.g.*, Kleinberg et al., *supra* note 14, at 240 (using these metrics of failure to appear and re-arrest to test the accuracy of bail hearing decisions).

102. *See* Corbett-Davies et al., *supra* note 14 ("The use of these algorithms often yields immediate and tangible benefits: Jail populations, for example, can decline without adversely affecting public safety." (citing studies in various jurisdictions)); Kleinberg et al., *supra* note 14, at 241; Jung et al., *supra* note 14.

103. *See, e.g.*, Dillon Reisman, *How New Jersey Used an Algorithm to Drastically Reduce Its Jail Population—and Why It Might Not Be the Right Tool for the Job*, ACLU N.J. (Aug. 30, 2022, 9:30 AM), https://www.aclu-nj.org/en/news/how-new-jersey-used-algorithm-drastically-reduce-its-jail-population-and-why-it-might-not-be [https://perma.cc/Q2AY-WLDT] (acknowledging New Jersey's pretrial risk assessment algorithm called the Public Safety Assessment (PSA) "drastically reduced its jail population at no cost to public

Second, some commentators contend that even if Step 1 is right, Step 2 is not. They argue that failing to appear at a court hearing is the wrong metric for "flight" because it does not distinguish between intentionally missing a court appearance by fleeing the jurisdiction and unintentionally missing a court appearance for "such mundane things as having a medical emergency, needing to care for a child, not having the day off from work, or simply forgetting."[104] They further argue that the correct metric for "flight" would assess the "probability that this person will intentionally evade prosecution," which is a measure that cannot be assessed with the available data.[105]

In a similar vein, some commentators argue that an arrest for a different crime is the wrong metric for "danger."[106] As one such commentator stated:

> Across this country, communities of color experience higher levels of police enforcement and surveillance activity than their white counterparts, which increases the likelihood and rate of arrest for people living in these communities of color, independent of whether more criminal activity actually exists there.[107]

---

safety" but criticizing it nonetheless because the "PSA has not addressed the systemic racial bias of who is being detained pending trial"). In addition, some judges might factor in the severity of the crime committed independent of the flight and danger risk of the arrested person. *See* Kleinberg et al., *supra* note 14, at 241.

104. Jeremy Cherson, *Policy Position Brief: On Pretrial Algorithms (Risk Assessments)*, THE BAIL PROJECT (July 15, 2022), https://bailproject.org/policy/pretrial-algorithms/ [https://perma.cc/P48M-7BVX]; *see also* Lauryn P. Gouldin, *Defining Flight Risk*, 85 U. CHI. L. REV. 677, 683 (2018) ("Flight risk is properly assigned to defendants who are expected to flee a jurisdiction. This is a small, and arguably shrinking, subcategory of a much larger group of defendants who pose risks of nonappearance.").

105. Cherson, *supra* note 104.

106. *See id.*

107. *See id.* Others have also commented on this issue. *See* Okidegbe, s*upra* note 99, at 2012 ("Arrest data, conviction data, and court appearance records are examples of the kinds of data these sources produce. It is well established that these sources produce data that are infected with racial and socioeconomic bias."); CHELSEA BARABAS ET AL., TECHNICAL FLAWS OF PRETRIAL RISK ASSESSMENTS RAISE GRAVE CONCERNS 1–2 (2019), https://pbtx.com/files/2019/12/TechnicalFlawsOfPretrial_ML_site.pdf ("[R]isk assessments frequently

These commentators further argue that the correct metric for "danger" would be "the probability that a person will commit a crime that results in harm to another person or persons while released pretrial," which is likewise a measure that cannot be assessed with the available data.[108]

Finally, some commentators argue that even if Step 1 and Step 2 are correct, Step 3 is not. These commentators argue that robots are no better than humans at predicting the risk that the arrested person will either fail to appear at the court hearing or be arrested before trial.[109]

The purpose of the above discussion is not to answer the question of whether robot judges are—or, with better programming, could be—more accurate than human judges in the bail hearing context. Rather, it is to first clearly delineate the steps necessary to make the "comparison path" argument and then demonstrate just how daunting the "comparison path" can be. Indeed, many areas of law, such as the later stages of criminal proceedings[110] and tort law,[111] will be eliminated at Step

---

define public safety risk as the probability of arrest. When tools conflate the likelihood of arrest for any reason with risk of violence, a large number of people will be labeled a threat to public safety without sufficient justification." (footnote omitted)).

108. Cherson, *supra* note 104.

109. *See, e.g.*, Julia Dressel & Hany Farid, *The Accuracy, Fairness, and Limits of Predicting Recidivism*, SCI. ADVANCES, Jan. 17, 2018, at 1, 1 (finding that "the widely used commercial risk assessment software COMPAS is no more accurate or fair than predictions made by people with little or no criminal justice expertise"); *see also* BARABAS ET AL., *supra* note 107, at 4 ("Pretrial risk assessments do not guarantee or even increase the likelihood of better pretrial outcomes.").

110. *See* Angelova et al., *supra* note 99, at 8 n.5 ("The single permissible objective of judges at the pretrial stage differs from later stages of the criminal justice system such as sentencing, where judges are permitted to consider multiple objectives at the same time . . . ."); *see also* KATE STITH & JOSÉ A. CABRANES, FEAR OF JUDGING: SENTENCING GUIDELINES IN THE FEDERAL COURTS 52 (1998) (discussing the "four generally recognized justifications for criminal penalties—retribution, deterrence, incapacitation, and rehabilitation" and noting "the tensions among these four rationales").

111. *See* RESTATEMENT (SECOND) OF TORTS § 901 (AM. L. INST. 1979) ("[T]he purposes for which actions of tort are maintainable. . . . are: (a) to give compensation, indemnity or restitution for harms; (b) to determine rights; (c) to

1 because there is no consensus as to the legal objective of that area of law. Even in cases where Step 1 can arguably be met, such as in contract law,[112] any colorable "comparison argument" will end at Step 2 because there will not be any conceivable metric to measure the ability of human judges or robot judges to achieve the Consensus on the Legal Objective. In short, there is an extremely limited number of legal areas where it can be plausibly argued using the "comparison path" that the robot judge has threaded the (three) needle(s) to demonstrate greater accuracy than the human judge.

However, there may be a way that the "comparison path" can have broader applicability—by taking a systematic approach rather than focusing on particular areas of law. For example, Eugene Volokh has argued that robot judges should replace human judges when their written opinions are found to be more persuasive than the written opinions of human judges, as determined by an outside panel of experts.[113] We can evaluate the strength of Volokh's proposal by using the three-step framework described above.

Step 1: Consensus on the Legal Objective

As discussed in the introduction, accuracy is the paramount objective of legal systems.[114] Volokh argues that it is not possible to determine the accuracy of legal decisions—because law is too indeterminate—but that we can determine a judicial opinion's value by its

---

punish wrongdoers and deter wrongful conduct; and (d) to vindicate parties and deter retaliation or violent and unlawful self-help."); Mark A. Geistfeld, *The Coherence of Compensation-Deterrence Theory in Tort Law*, 61 DEPAUL L. REV. 383, 384 (2012) (noting that "there is no consensus about the appropriate rationale for tort liability").

112. *See* Hoffman & Arbel, *supra* note 70, at 461 (stating that there is a legal consensus that contract interpretation has a common objective—ascertaining the parties' intention at the time the parties made their contract—and noting that, in theory, this question has a correct answer) (citing Alan Schwartz & Robert E. Scott, *Contract Theory and the Limits of Contract Law*, 113 YALE L.J. 541, 568 (2003)).

113. *See* Volokh, *supra* note 8, at 1152.

114. *See supra* Introduction.

"persuasiveness" as a stand-in for accuracy.[115] As Part I argued, I disagree with the proposition that the law is always indeterminate.[116] Moreover, I contend that persuasion cannot be the objective of a legal opinion because a legal opinion can be "persuasive" but wrong and not "persuasive" but right. For example, let's return to the "No Vehicles in the Park" hypothetical and assume that the case is about a father with a baby stroller. Let's further assume that you initially strongly believe that the father should not be issued a citation. However, after reading a very "persuasive" opinion that the father should be issued a citation, you still believe that the father should not be issued a citation, but you view the case as a fairly close call. Alternatively, after reading a one-line cursory opinion that the father should not be issued a citation, you do not change your initial position at all.

In short, you found the first opinion more persuasive than the second—in the sense that it moved you further away from your initial position—but you still consider it wrong, and you found the second opinion not at all persuasive, but you still consider it right. Accordingly, greater "persuasiveness" cannot serve as the consensus legal objective for judicial decisions because it does not always correspond with greater accuracy.

Step 2: Specified Metric

Volokh argues that the specified metric to measure the ability of human judges or robot judges to achieve the Consensus on the Legal Objective (in this case,

115. *See* Volokh, *supra* note 8, at 1153, 1192. Volokh argues that "[t]he quality [of legal decisions] should largely be measured using the metric of persuasiveness" which eliminates the "need to decide what *the* supposedly correct answer is." *See id.* at 1153, 1192. In making this argument, Volokh equates two questions as being essentially the same: "Did the opinion persuade me?" and "Did it lead me to conclude that the result is legally correct, however I might understand 'correctness' for this particular legal question?" *See id.* at 1152. As explained in the text of this article, I disagree that these questions are essentially the same.

116. *See supra* Part I.

persuasiveness) should be an outside panel of blue-ribbon experts, such as retired judges.[117] However, America's history of "blue ribbon" panels, particularly in areas such as law, reveals that there often is no consensus that the selected experts are the actual "experts."[118] Without such a consensus, outside experts cannot be used to measure persuasiveness.

Step 3: Empirical Demonstration

This step would utilize an outside panel of experts to demonstrate empirically that the written opinions of robot judges are considered more persuasive than the opinions of human judges. We do not get to Step 3 because of the problems with Steps 1 and 2.

I propose that a better systematic approach for the "comparison path" is to keep accuracy, without modification, as the consensus legal objective and then use "rate of agreement with human judges who are higher in the judicial hierarchy"—or phrased conversely, the "reversal rate"—as the specified metric.[119] This

117. *See* Volokh, *supra* note 8, at 1154.

118. *See* Will Rhee & Claire Flynn Sellers, *Retooling Blue-Ribbon Advisory Committees for a Post-Fact World*, 125 W. VA. L. REV. 451, 456 (2022) (discussing American skepticism of expert commissions and the "populist backlash against the 'best and the brightest' that the blue-ribbon represents"); *see also Tough Policy Questions Often Subject to Death by "Blue Ribbon" Commission*, CBS NEWS (Mar. 13, 2018, 7:03 AM), https://www.cbsnews.com/news/tough-policy-questions-often-subject-to-death-by-blue-ribbon-commission/ [https://perma.cc/A9XH-DL6T] (noting that the members of expert commissions "are not elected or accountable to the public" and there is "no quality control").

119. The closest proposal to mine in the literature is Engtron and Ho's "prospective benchmaking" proposal. *See generally* David Freeman Engtron & Daniel E. Ho, *Algorithmic Accountability in the Administrative State*, 37 YALE J. ON REG. 800 (2020) (proposing prospective benchmarking as a tool to provide oversight for AI tools used by the government). "The core idea [of this proposal] is that when agencies adopt an AI decision making tool, they should subject it to *benchmarking* relative to a random hold-out set of cases that undergo conventional human review." *Id.* at 849. In other words, Engtron and Ho's proposal calls for the accuracy of robot judges to be tested by comparing the outputs of the robot judges to the outputs of human judges on the s*ame* rung of the hierarchy, with the implicit assumption that the human judges are more accurate than the robot judges and the question is whether the robot judges are "accurate enough" to justify their use based on efficiency. My proposal is

proposal takes advantage of the fact that it is implicit in hierarchical judicial systems that decisions of the judges higher in the hierarchy are more accurate than the decisions of the judges lower in the hierarchy.[120] Accordingly, if robot judges and human judges are on the same level of the hierarchy, we can measure the accuracy of their decisions by the agreement rate with human judges higher up in the hierarchy.[121] Notably, prominent jurists have already suggested that a judge's reversal rate can be a valid indicator of a judge's performance.[122]

Implementing my proposal in the typical three-tiered judicial hierarchy in the U.S. federal and state courts (trial court, intermediate appellate court, and high court) will not be straightforward. First, some judicial decision making involves making factual determinations based on credibility judgments arising out of the observation of live witness testimony, a task

---

different because it does not assume that human judges are more accurate than robot judges and provides an opportunity for robot judges to demonstrate superior accuracy through greater agreement with human judges situated on a *higher* rung of the judicial hierarchy.

120. *See* JAMES E. PFANDER, ONE SUPREME COURT: SUPREMACY, INFERIORITY, AND THE JUDICIAL DEPARTMENT OF THE UNITED STATES 41 (2009) (arguing that "the Framers' very conception of a unitary and hierarchical, rather than a plural and horizontal, judiciary presupposed a duty on the part of lower courts to obey their superior"). Indeed, as Justice Jackson famously wrote about the U.S. Supreme Court, "We are not final because we are infallible, but we are infallible only because we are final." Brown v. Allen, 344 U.S. 443, 540 (1953) (Jackson, J., concurring). In other words, the U.S. judicial system assumes that the decisions of the judges at the top of the judicial hierarchy are "correct" regardless of their "correctness" in an abstract, philosophical sense.

121. While there is some debate in the literature whether, in the absence of a direct precedent, lower courts should use their own judgment or predict what the court higher in the judicial hierarchy will do, the so-called predictive approach offers the advantages of judicial economy and consistency of interpretation. *See generally* Evan H. Caminker, *Precedent and Prediction: The Forward-Looking Aspects of Inferior Court Decisionmaking*, 73 TEX. L. REV. 1 (1994) (discussing both approaches and finding the predictive approach superior for these reasons).

122. *See* Richard A. Posner, *Judicial Behavior and Performance: An Economic Approach*, 32 FLA. ST. U. L. REV. 1259, 1259 (2005) (writing that a judge's reversal rate can be "a critical performance criterion"). In a similar vein, Chief Justice Roberts stated at a recent oral argument: "I wonder if I'm a court of appeals judge, it seems to me the most important thing in deciding the case is to make sure that I'm not reversed." Transcript of Oral Argument at 9, Hughes v. United States, 584 U.S. 675 (2018) (No. 17-155).

that would appear to be ill-suited for robot judges. Indeed, in his 2023 Year-End Report on the Federal Judiciary, Chief Justice John Roberts emphasized this point in suggesting that robot judges could never "fully replace" human judges:

> Judges . . . measure the sincerity of a defendant's allocution at sentencing. Nuance matters: Much can turn on a shaking hand, a quivering voice, a change of inflection, a bead of sweat, a moment's hesitation, a fleeting break in eye contact. And most people still trust humans more than machines to perceive and draw the right inferences from these clues.[123]

Leaving aside questions about the appropriateness of basing sentencing decisions on "a fleeting break in eye contact," I agree with Chief Justice Roberts's general point that robot judges are incapable (as of now) of deciding cases at the trial level that require (or are perceived to require) making such credibility judgments.[124] Moreover, even if robot judges were capable, it would not be possible to test the correctness of such judgments using the comparison approach because the robot trial judges would be making their decisions based on information that would not be accessible to the human appellate judges.

---

123. ROBERTS, *supra* note 8, at 6.

124. However, it certainly appears possible that as the technology develops, a robot judge that solely analyzes the written testimony of witnesses may be more accurate in making credibility judgments than a human judge or jury who observes the live testimony. *See* Kiel Brennan-Marquez & Julia Ann Simon-Kerr, *Judging Demeanor*, 109 MINN. L. REV. (forthcoming) (manuscript at 3), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4776770 (arguing that observations of defendant demeanor by jurors often leads to less accurate fact-finding as it is "opens the door to bias, stereotyping, and cultural chauvinism"); Riccardo Loconte et al., *Verbal Lie Detection Using Large Language Models*, 13 SCI. REPS. 22849 (2023) (finding in a laboratory experiment that an LLM was more accurate than humans at detecting whether written narratives, supplied by human volunteers, were true or false). The issue here is the metric to test the accuracy of robot judges vs. human judges. One possibility is that robot judges could be tested on transcripts of trial testimony where subsequent evidence (e.g., DNA testing) later confirmed or contradicted the accuracy of the fact-finding by the human decision maker(s). In any event, this would be a topic for a different article.

However, many decisions made by trial judges, such as summary judgment motions and motions to dismiss,[125] are based solely on a written record. So long as cases are randomly assigned to the trial judges without regard to whether the judge is a robot or human, my proposal could be used to analyze human judges and robot judges at the trial level for these types of decisions, using the rate of agreement with human judges at the intermediate appellate level as the specified metric. Furthermore, since a losing litigant is more likely to be willing to incur the cost of an appeal if the trial court decision is less accurate,[126] so long as cases are randomly assigned to the trial judges without regard to whether the judge is a robot or human, the rate of appeal of trial judge decisions could be used as another specified metric. Indeed, this can be viewed as analogous to how "the arguments basically stopped" after minor league baseball implemented the robot umpire.[127]

My proposal could also be used to analyze human judges and robot judges on intermediate appellate courts using the rate of agreement with human judges on the high court. In his 2023 Year-End Report on the Federal Judiciary, Chief Justice Roberts contends that "[a]ppellate judges, too, perform quintessentially human functions" as "[m]any appellate decisions turn on whether a lower court has abused its discretion, a standard that by its nature involves fact-specific gray areas."[128] However, because appellate judges are simply reviewing a written record—and are not making credibility decisions based on witness testimony—it is unclear why robot appellate judges could not perform this task as well as human judges.

---

125. For example, in motions to dismiss for failure to state a claim made under Rule 12(b)(6) of the Federal Rules of Civil Procedure, the non-conclusory facts alleged in the complaint are assumed to be true. *See* Ashcroft v. Iqbal, 556 U.S. 662, 664 (2009).

126. *See generally* Heise & Wells, *supra* note 51 (comparing a plaintiff's trial court success with preserving that success on appeal).

127. *See supra* note 64 and accompanying text.

128. *See* ROBERTS, *supra* note 8, at 6.

Chief Justice Roberts also argues that robot judges cannot perform appellate decision making that "focus[es] on open questions about how the law should develop in new areas" because "AI is based largely on existing information, which can inform but not make such decisions."[129] However, the various possible paths that the law can take are generally well understood and, if there is a "hidden" path to be discerned, it is an open question whether human or robot judges are more likely to find it.[130]

In any event, to the extent that Chief Justice Roberts is correct that robot judges are incapable of making accurate appellate decisions based on ambiguous facts or open questions of law, then human judges higher up in the judicial hierarchy will presumably find robot judges to be less accurate than human judges. But if robot appellate judges are more likely to match the decisions of human judges higher up in the judicial hierarchy, it would be difficult to argue that they cannot handle these types of cases.

However, other problems remain. One problem is that because high courts generally have discretionary review—and in practice only agree to hear a small fraction of the cases in which a review is sought—it would be difficult to draw any conclusions about an intermediate appellate judge's overall accuracy from these unusual cases.[131] This issue can be resolved in two ways. First, the high court can select on a random basis a certain number of cases from the intermediate

129. *See id*.

130. For example, Adam Unikowsky tested whether an LLM (Claude 3 Opus) could generate legal standards for recently decided U.S. Supreme Court cases that neither of the parties proposed in its briefs. *See* Adam Unikowsky, *In AI We Trust, Part II: Wherein AI Adjudicates Every Supreme Court Case*, ADAM'S LEGAL NEWSL. (June 16, 2024), https://adamunikowsky.substack.com/p/in-ai-we-trust-part-ii [https://perma.cc/X9HN-DBBG]. He found that the LLM could "generate novel legal standards" which on at least one occasion were "clearer and more administrable than anything proposed by the parties or the [U.S. Supreme] Court." *Id.*

131. *See* Ryan W. Copus, *Statistical Precedent: Allocating Judicial Attention*, 73 VAND. L. REV. 605, 608 (2020) (noting that the U.S. Supreme Court reviews only 0.1% of circuit court decisions).

appellate courts for automatic review. Second, legislatures could create a new tier between the intermediate appellate courts and high court that would automatically review a certain number of cases from the intermediate appellate courts on a random basis. Interestingly, proposals to have the high court review some cases on a random basis or to create a new tier between the intermediate appellate court and the high court have already been made to increase the accountability and consistency of the U.S. legal system, among other reasons.[132] The ability to test the accuracy of robot judges would be another reason to adopt these proposals.

A second problem is that there may be times where the accurate decision at a lower tier of the judicial hierarchy differs from the accurate decision at a higher tier. This can occur when the lower court follows a pre-existing direct precedent and the higher court explicitly overrules that precedent in that very case.[133] The easiest solution to this problem is to exclude such cases, which are undoubtedly small in number, from the comparison set.[134] However, I would not exclude cases where some commentators believe that a higher court ignored or distinguished an on-point precedent because the higher

132. *See* Daniel Epps & William Ortman, *The Lottery Docket*, 116 MICH. L. REV. 705, 707–08 (2017) (proposing that the U.S. Supreme Court "supplement the traditional certiorari docket with a small number of cases randomly selected from final judgments of the circuit courts" and arguing that their proposal would provide critical information to the U.S. Supreme Court and increase the accountability of the U.S. Courts of Appeals); Michael Abramowicz, *En Banc Revisited*, 100 COLUM. L. REV. 1600, 1611–12 (2000) (discussing proposals to add a new tier between the U.S. Supreme Court and the U.S. Courts of Appeals in order to resolve circuit splits and alleviate caseload pressure on the U.S. Supreme Court).

133. *See, e.g.*, Rodriguez de Quijas v. Shearson/Am. Express, Inc., 490 U.S. 477, 484 (1989) ("If a precedent of this Court has direct application in a case, yet appears to rest on reasons rejected in some other line of decisions, the Court of Appeals should follow the case which directly controls, leaving to this Court the prerogative of overruling its own decisions.").

134. These cases will also be readily identifiable because they are, by definition, only cases where the higher court explicitly overruled a prior precedent.

court is ultimately the decision maker as to the applicability of the precedent.

It is instructive to apply the concepts discussed above to a particular court system to see how they might work in practice. The U.S. Immigration Court System provides a good example because its features—intermediate appellate review by a single judge, high court review that is not discretionary, and a relatively simple process to enact structural change—do not raise certain issues that may be present in other court systems. To be clear, I am not advocating for any particular reform of the U.S. Immigration Court System, and before robot judges are used in any system, such use would first need to be tested and refined in practice or lower stake settings.

By way of background, the lowest level of the U.S. Immigration Court System consists of approximately 650 immigration judges who act as trial judges and decide questions of both fact and law, with each judge adjudicating approximately 200 to 500 cases per year.[135] The intermediate appellate level consists of 23 Board of Immigration Appeals ("BIA") judges who decide appeals based solely on the written record from the lower court.[136] About one-fifth of the cases immigration judges decide are appealed to the BIA judges.[137] Appeals are generally decided by a single BIA judge without a written decision.[138] In deciding these appeals, the BIA judges give deference to the factual determinations of the immigration judges but not to their legal determinations.[139] Decisions from BIA judges can then

135. *See* MUZAFFAR CHISHTI ET AL., MIGRATION POL'Y INST., AT THE BREAKING POINT: RETHINKING THE U.S. IMMIGRATION COURT SYSTEM 8, 16 (2023), https://www.migrationpolicy.org/sites/default/files/publications/mpi-courts-report-2023_final.pdf [https://perma.cc/FM8J-FGR8].

136. *See id.* at 8; 8 C.F.R. § 1003.1(d)(3)(iv) ("The [BIA] will not engage in factfinding in the course of deciding cases.").

137. *See* CHISHTI ET AL., *supra* note 135, at 17.

138. *See* Faiza W. Sayed, *The Immigration Shadow Docket*, 117 NW. U.L. REV. 893, 897 (2023).

139. *See* ANDREW PATTERSON ET AL., AM. IMMIGR. COUNCIL, STANDARDS OF REVIEW APPLIED BY THE BOARD OF IMMIGRATION APPEALS 2 (2020),

be appealed to a U.S. Court of Appeals, which happens in about one in four cases.[140] Accordingly, the structure of the U.S. Immigration Court follows the typical three-tiered structure of other U.S. legal systems, but with a key difference: the "high" court (here, a U.S. Court of Appeals) does not have discretionary review and is obligated to hear every case that is appealed for which it has jurisdiction.[141] As a result, robot judges can be incorporated—and tested against human judges—at the BIA level, with both the rate of agreement with a U.S. Court of Appeals and the rate of appeal to a U.S. Court of Appeals as the specified metrics.[142]

An alternate approach would be to adjust the structure of the U.S. Immigration Court system so the "rate of agreement" test could be applied to a random set of cases instead of the set of cases that are appealed to a U.S. Court of Appeals. This alternate approach could enhance the "rate of agreement" test because cases that are appealed to a U.S. Court of Appeals may have special characteristics that skew the results. To implement the alternate approach, a subset of the 23 BIA judges could be promoted to a newly created level between the BIA and the U.S. Courts of Appeals. Let's call it the

---

https://www.americanimmigrationcouncil.org/sites/default/files/practice_advisory/standards_of_review_applied_by_the_board_of_immigration_appeals.pdf [https://perma.cc/P9TN-REWM].

140. *See* CHISHTI ET AL., *supra* note 135, at 18.

141. *See* Nasrallah v. Barr, 590 U.S. 573, 580 (2020) ("[A] noncitizen's various challenges arising from the removal proceeding must be 'consolidated in a petition for review and considered by the courts of appeals.'" (quoting INS v. St. Cyr, 533 U.S. 289, 313, and n.37 (2001))). There are some immigration appeals that are excluded from the jurisdiction of the U.S. Courts of Appeals, including those raising purely questions of fact. *See*, *e.g.*, Patel v. Garland, 596 U.S. 328 (2022) (holding that federal courts lack jurisdiction to review facts raised in an application for discretionary relief from removal). These cases can be excluded from the set of cases in which the comparison between the robot judges and human judges at the BIA level are made.

142. Indeed, some scholars have argued that the high rate of appeals from BIA judge decisions to a U.S. Court of Appeals suggests a high error rate of BIA decisions. *See* CHISHTI ET AL., *supra* note 135, at 18 (arguing that "[t]he disproportionate representation of immigration court cases among administrative agency appeals to the federal circuit courts . . . lends credence to concerns about the quality of decision-making in immigration courts").

Reviewing Immigration Court. The judicial openings at the BIA level created by the promoted BIA judges would then be filled with robot judges. The Reviewing Immigration Court would sit en banc and automatically review a random set of cases from the robot and human BIA judges.[143] Accordingly, this alternate approach would provide a true apples-to-apples test between the human judges and the robot judges at the BIA level, with the rate of agreement with the Reviewing Immigration Court as the specified metric. While making a structural change to a court system might be a "big lift" in most circumstances—raising constitutional issues or at least requiring legislative authorization—here it would appear that a U.S. Attorney General could make this structural change on his or her own.[144]

Moreover, using either of my proposed approaches, appellate robot judges could very well prove to be more accurate than appellate human judges at the BIA level, as BIA judges have long been criticized for their low "quality"[145] and "error-prone"[146] decision making. One U.S. Court of Appeals decision noted that in the past year "different panels of this court reversed the Board of Immigration Appeals in whole or part in a staggering 40 percent" of the cases and that the "adjudication of these cases at the administrative level has fallen below the minimum standards of legal justice."[147] A 2023 article that reviewed U.S. Courts of Appeals decisions concluded that "[o]bvious errors in BIA decision-making persist to this day."[148] Indeed, similarly to how human umpires

143. Under this proposed structure, the decisions of the BIA judges that are not reviewed by the Reviewing Immigration Court and the decisions of the Reviewing Immigration Court could be appealed to a U.S. Court of Appeals.

144. *See* Sayed, *supra* note 138, at 902 (noting that by the pertinent statute, "[BIA] members are merely 'attorneys appointed by the Attorney General to act as the Attorney General's delegates in the cases that come before them'" (quoting 8 C.F.R. § 1003.1(a)(1) (2022) (amended 2024))).

145. *See* CHISHTI ET AL., *supra* note 135, at 17 (noting that the "[f]ederal courts of appeal, scholarly articles, and congressional studies have strongly criticized the quality of immigration court decisions").

146. *See* Sayed, s*upra* note 138, at 942.

147. Benslimane v. Gonzales, 430 F.3d 828, 829–30 (7th Cir. 2005).

148. Sayed, *supra* note 138, at 943.

make clear-cut errors because the speed of the pitch may be too fast for them to accurately process, many have speculated that the cause of the errors by BIA judges is that they are required to decide cases—estimated at five per business day—at a rate that is simply faster than the rate at which human judges can accurately function.[149]

Finally, to the extent that robot BIA judges can demonstrate superior accuracy over the human BIA judges, as determined using either of my proposed approaches, more robot judges could be added to the BIA level, either in addition to or in replacement of some of the human BIA judges. To the extent that the addition of more robot BIA judges reduces the workload of the human BIA judges, adding robot BIA judges would have the added benefit of potentially increasing the accuracy of the human BIA judges. A certain number of human judges should remain at the BIA level, however, in order to continually test that the accuracy rate of the robot BIA judges remains superior to the accuracy rate of the human BIA judges.

III. ACCURACY BY PROCESS

A. Majority Decision Making and the Condorcet Jury Theorem

The third path for robot judges to demonstrate superior accuracy is arguably the path of least resistance, which I call majority robot decision making. In majority robot decision making, the robot judge

149. *See id.* at 944 ("[T]he sheer number of appeals, the small number of Board members, and the streamlining reforms may explain the inconsistency in decisions, as well as the low quality of Board decisions and frequent legal errors committed by the Board. Since 2008, the BIA has received nearly 30,000 appeals each year. During this time, the number of Board members has fluctuated, but it has never been more than twenty-three members, the maximum permitted by current regulations. This means that at a minimum (using 30,000 decisions and twenty-three members and assuming single member opinions), each member is responsible for deciding 1,304 appeals a year."); *see also Benslimane*, 430 F.3d at 830 (speculating that the cause of the BIA's and Immigration Court's legal errors might be due to "resource constraints").

replicates the decision of the majority of qualified human judges. I write that this is the "path of least resistance" because it is generally assumed that a robot judge will "generate" its decisions based on "past human practice,"[150] and majority robot decision making is a variant of this amorphous category. However, I argue that majority robot decision making is the only variant that has a claim to being more accurate than human judges.

The theoretical basis for majority robot decision making is the Condorcet Jury Theorem, which demonstrates mathematically that under certain conditions, "the more people we ask a question, the greater the chance the majority of them will select the correct answer."[151] The Condorcet Jury Theorem is particularly promising in the legal setting because one of its core assumptions—that each member of the group has a better-than-random chance at arriving at the correct answer—would appear to be satisfied as judges are specifically selected for their legal expertise.[152]

---

150. *See* Anthony J. Casey & Anthony Niblett, *Will Robot Judges Change Litigation and Settlement Outcomes?*, MIT COMPUTATIONAL L. REP. (Aug. 14, 2020), https://law.mit.edu/pub/willrobotjudgeschangelitigationandsettlementoutcomes/release/1#:~:text=Here%2C%20we%20show%20that%20the,to%20translate%20predictions%20into%20judgments [https://perma.cc/SW8V-L5MX] (arguing that the most promising way for automating judicial decisions is via "predictive litigation algorithms that use historical data from prior case outcomes to determine how the new case fits within the contours of existing law").

151. *Fitzpatrick*, *supra* note 49, at 111; *see also* Ruth Ben-Yashar & Mor Zahavi, *The Condorcet Jury Theorem and Extension of the Franchise with Rationally Ignorant Voters*, 148 PUB. CHOICE 435, 435 (2011) ("In the simplest form of the [Condorcet Jury Theorem], a group of decision makers votes independently on a binary choice with each voter having the same probability $p > 1/2$ of choosing the correct alternative. The Theorem indicates that, the larger is the group, the better the group performs in terms of the likelihood of making the correct decision by majority voting.").

152. As Michael Abramowicz argues, if you disagree with this notion, the entire enterprise of judging would essentially be a "waste of time" as judges could essentially be replaced with coin flips. *See* Abramowicz, *supra* note 132, at 1630 n.123. Moreover, you would need to be ambivalent as to whether, for example, in a U.S. Supreme Court decision that split 8 to 1, the majority opinion of eight Justices or the dissenting opinion of one Justice became the law of the land. *See id.* at 1631 n.125.

Indeed, most legal systems provide that where decisions are made by multiple judges, the decisions are made by majority vote, which arguably is an implicit endorsement of the Condorcet Jury Theorem.[153]

However, four preconditions need to be established before majority robot decision making can be justified by the original formulation of the Condorcet Jury Theorem:

1. The legal decision must be binary, i.e., a choice between two alternatives;
2. There must be a defined set of human judges who are deemed equally qualified to hear the case;
3. Such judges must arrive at their decisions independently; and
4. Each such judge is equally likely to be assigned to the case.

Once these preconditions are met, a robot judge engaging in majority robot decision making would predict how each of the human judges would rule on the matter at hand and issue the decision that the majority of human judges in the defined set would issue, which would also be reflective of the median vote. Thus, a key advantage of majority robot decision making is that it addresses the "black box" problem by providing an explanation about how the robot judge reached its decision.

One complication with this process is that the predictions of robot judges are generally based on probabilities, i.e., human judge *A* has a 70% chance of granting the motion, not human judge *A* will grant the motion. Accordingly, there are (at least) two ways to determine the "majority decision" of the defined set of

153. *See* Guha Krishnamurthi, *For Judicial Majoritarianism*, 22 U. PA. J. CONST. L. 1201, 1204 (2020) ("[W]hen a court is comprised of multiple judges it must also have a procedure for rendering a decision when judges disagree, and most courts use majority vote."); Jeremy Waldron, *Five to Four: Why Do Bare Majorities Rule on Courts?*, 123 YALE L.J. 1692, 1695 (2014) ("[Majority decision] is pretty much universal among multi-member judicial panels . . . .").

human judges: what I call the "50% Plus" approach and the "Vegas Oddsmaker" approach.[154]

The following example will illustrate the two approaches. Let's suppose the defined set consists of three human judges, A, B, and C. Let's further suppose that the robot judge predicts that there is a two out of three chance that judges A and B will each grant the motion while there is a 100% chance that judge C will deny the motion. If you simply predict how each judge is going to vote in isolation, i.e., assign the judge's vote when the prediction crosses the 50% threshold, there would be two votes for granting the motion and one vote for denying the motion. Under the 50% Plus approach, the majority decision would therefore be to grant the motion.

However, if a Vegas Oddsmaker were setting the odds on whether the motion would be granted, the denial of the motion would be favored, assuming that a particular human judge (A, B, or C) was going to be randomly assigned to the case.[155] In theory at least, the Vegas Oddsmaker approach is the preferable way to determine the majority opinion in the defined set because it better captures the "true" majority opinion. In practice, however, the Vegas Oddsmaker approach may give too much sway to the more predictable judges.

Assuming this issue gets resolved, implementing the "accuracy by process" path based on the Condorcet Jury Theorem is relatively straightforward. For example, let's imagine a "robot" as a federal district court judge in a particular U.S. federal court district deciding a defendant's motion to dismiss a complaint for failure to state a claim pursuant to Rule 12(b)(6) of the Federal Rules of Civil Procedure. Here, the four preconditions

---

154. Casey and Niblett were the first to raise the issue of how to translate "robot judge probability predictions" into automated robot judicial decisions, arguing that using a 50% Plus approach would distort settlement outcomes. *See* Casey & Niblett, *supra* note 150.

155. In this example, the likelihood of the motion being "denied" is five out of nine, i.e., the likelihood of each judge being randomly assigned to the case multiplied by the likelihood of such judge denying the motion ((1/3 x 1/3) + (1/3 x 1/3) + (1/3 x 1) = 5/9).

discussed above for the Condorcet Jury Theorem are met. Precondition number one is satisfied because the decision is (essentially) binary—either the complaint is dismissed or it is not.[156] Precondition number two is satisfied because there is a defined set of human judges who are deemed equally qualified to hear the case, i.e., any of the U.S. federal district court judges in that federal court district. Precondition number three is satisfied because it is a single-judge deliberation. Precondition number four is satisfied because any of these judges could be randomly assigned to hear the case. Moreover, the process necessary for majority robot decision making appears to be feasible based on the technology as it exists today.[157] For example, a private-sector service called Pre/Dicta already makes predictions about how each U.S. federal district court judge will decide motions to dismiss based on how the judge has ruled in prior cases as well as the background of the judge.[158]

It is instructive to consider whether the U.S. Federal Sentencing Guidelines, with some modifications, could be considered an early version of robot decision making under this path because the Guidelines were purportedly "primarily" based upon "typical, or average, actual past practice."[159] If the Guidelines specified a non-

156. In actuality, a judge could deny a Rule 12(b)(6) motion with leave to amend, which is a compromise decision between a complete grant and a complete denial. However, any denial of a Rule 12(b)(6) motion by a robot judge could automatically be with leave to amend, making the decision completely binary.

157. *See* Casey & Niblett, *supra* note 150; Chen et al., *supra* note 6, at 132 nn.23–24 (listing the various models that researchers have developed to predict case outcomes at the U.S. Supreme Court and other tribunals).

158. *See* Cassandre Coyer, *Pre/Dicta Expands Litigation Prediction Platform with New Motion Types, Case Timelines*, LAW.COM (Nov. 14, 2023, 9:00 AM), https://www.law.com/legaltechnews/2023/11/14/predicta-expands-litigation-prediction-platform-with-new-motion-types-case-timelines/ [https://perma.cc/RE7H-KJVV].

159. Stephen Breyer, *The Federal Sentencing Guidelines and the Key Compromises upon Which They Rest*, 17 HOFSTRA L. REV. 1, 7, 17 (1988) ("[I]n creating categories and determining sentence lengths, the Commission, by and large, followed typical past practice, determined by an analysis of 10,000 actual cases."). *But see* NAT'L RSCH. COUNCIL, THE GROWTH OF INCARCERATION IN THE UNITED STATES: EXPLORING CAUSES AND CONSEQUENCES 78 (2014) (finding

discretionary fixed number that reflected the median sentence, and not a discretionary range as they do in their current form,[160] then sentencing under the Guidelines might be considered more accurate than pre-Guidelines practice under an expansive view of the Condorcet Jury Theorem.[161] However, even with this modification, the Guidelines cannot be completely justified under this path because they "froze" typical practice at the time of their creation. While they might have reflected the median view of U.S. federal district court judges at one point in time, there is no assurance that they reflect the median view going forward.

We can imagine further modifications to the Guidelines to perhaps fix this problem. For example, now that that the Guidelines are advisory, perhaps they can be automatically "updated" each year to reflect the current median sentencing practice of U.S. federal district court judges.[162] However, even advisory Guidelines are going to inevitably affect the decision making of at least some judges,[163] particularly since judges are instructed to use the advisory range as a

---

that the federal sentencing guidelines "greatly increased both the percentage of individuals receiving prison sentences and the length of sentences for many offenses").

160. *See generally* United States v. Booker, 543 U.S. 220 (2005) (effectively changing the federal sentencing guidelines from a mandatory range to an advisory range).

161. *See* Cass R. Sunstein, *Group Judgments: Statistical Means, Deliberation, and Information Markets*, 80 N.Y.U. L. REV. 962, 974 (2005) (arguing that, in the context of non-binary, numerical decisions, "[t]he accuracy of the median judgment, for large groups, is simply an application of the Condorcet Jury Theorem").

162. *See generally* Mark H. Allenbaugh, *Sentencing in Chaos: How Statistics Can Harmonize the "Discordant Symphony,"* 32 FED. SENT'G REP. 128 (2020) (arguing that the sentencing ranges in the Guidelines should be recalibrated downward to reflect current sentencing practice).

163. *See* Crystal S. Yang, *Have Interjudge Sentencing Disparities Increased in an Advisory Guidelines Regime? Evidence from* Booker, 89 N.Y.U. L REV. 1268, 1286–87 (2014) (finding that after the Guidelines became advisory, "[s]ome courts sentenced with minimal consideration of the applicable Guidelines range, while others treated the Guidelines as a dominant factor").

"starting point"[164] and sentences that fall within the advisory range are less likely to be reversed,[165] so it will be impossible to capture what would have been the "true" median sentence in the absence of the Guidelines.

After the Guidelines became advisory, federal district court judges have been much more likely to use their discretion to issue sentences below the range than above it.[166] However, these practices do not necessarily mean that the Guidelines are causing the median sentence to be more severe than what it would be in the absence of the Guidelines. It could just mean that judges who would sentence more severely in the absence of the Guidelines are more likely to follow the Guidelines than judges who would sentence more leniently.[167]

In any event, under the "accuracy by process" approach I am presenting here, there cannot be anything, such as the Guidelines, that has any degree of coercive power over the sentencing decisions of the federal district court judges, because that would violate the independent judgment pre-condition of the

---

164. *See* Gall v. United States, 552 U.S. 38, 49 (2007) ("As a matter of administration and to secure nationwide consistency, the Guidelines should be the starting point and initial benchmark.").

165. *See* Rita v. United States, 551 U.S. 338, 341 (2007) (holding that the U.S. Courts of Appeals may presume that sentences within the Guidelines range are reasonable).

166. *See* Kimberly Kaiser & Cassia Spohn, *Why Do Judges Depart? A Review of Reasons for Judicial Departures in Federal Sentencing*, 19 CRIMINOLOGY, CRIM. JUST. L. & SOC'Y 44, 48 (2018). For instance, in 2013, excluding government-sponsored sentences (which usually occur when the prosecutor requests a downward departure based on the defendant's cooperation), there were 16,421 sentences outside the Guidelines range, 14,740 of which were below the range and just 1,681 of which were above the range. *Id.*; *see also* Robert J. Anello & Richard F. Albert, *Life After "Booker": Insights from Federal Sentencing Data*, N.Y.L.J. (Aug. 15, 2018) ("Focusing on the all-important category of nongovernment-sponsored reductions, the proportion of such sentences nationwide has shown a gradually increasing trend over the post-*Booker* period, from 12.1% in 2006 to 17.4% in 2011 to 20.1% in 2017, with some leveling off in more recent years.").

167. This may very well be the case. *See* Sam J. Merchant, *A World Without Federal Sentencing Guidelines*, 102 WASH. U. L. REV. (forthcoming 2025) (manuscript at 1) (suggesting that federal district judges would sentence more severely in the absence of the Guidelines based upon a study of sentencing practices for offenses that are not covered by a direct guideline).

Condorcet Jury Theorem.[168] We would need to give the human sentencing judges maximum discretion in order for the robot sentencing judges to achieve maximum accuracy. I suspect that in such a world, the decision making of the human judges (with all the disparities that will inevitably occur) will prove to be more controversial than the decision making of the robot judges.

B. Mixed Ideological Appellate Panels Combined With a "Culture" of Compromise

While the Condorcet Jury Theorem can readily supply a justification for majority robot decision making at the single-judge federal district court level, the justification becomes more tenuous at the multi-judge appellate levels. The problems are at least threefold. First, if a subset panel is used, such as the usual three-judge panel at the U.S. Courts of Appeals, there is no assurance that the majority vote of the panel would replicate the majority vote of all the judges of the court.[169] Second, if an en banc panel is used with a large number of judges, the "independent" prong of the Condorcet Jury Theorem is less likely to be satisfied for reasons explained below in the discussion of panel size. Third, the important question in appellate decisions is often not the binary question of which party wins but the non-binary question of where on the ideological spectrum the precedent is set.

However, there may be a process-based argument that does not need to be strictly tethered to the Condorcet Jury Theorem. If: (1) a consensus exists among the relevant stakeholders that a particular process is more likely to lead to accurate decisions; and (2) the robot judges are able to adhere to that process and the human

168. However, sentencing statistics could be provided to the judges.

169. S*ee* Hon. Albert Branson Maris, Hearing and Rehearing Cases in Banc, 14 F.R.D. 91, 96 (1954) ("A decision of a controversial question made by a majority of all the judges of the court in banc obviously has much greater authority than a decision by two concurring judges of a panel of three which all the other five judges of the court might consider quite erroneous.").

judges are not, then an argument can be made that the robot judge decisions are more accurate. Although I do not contend that there is currently any process for which such a consensus exists, I will propose a candidate for such a process in the multi-judge, appellate decision making context.

My argument for this candidate will be brief because the purpose of this discussion is not to create a consensus for this candidate but rather to show how robot judges can potentially adhere to a "consensus process" if one forms and the technology allows for it.

I contend that multi-judge, appellate decision making is more likely to result in accurate decisions when the process has the following three characteristics: (1) the judges on the panel are ideologically diverse; (2) the panel is small in size; and (3) the panel makes its decision under a majority vote rule but with a significant degree of pressure on the judges to reach a unanimous decision. I will now briefly argue in broad strokes why these characteristics are the hallmarks of accurate multi-judge, appellate decision making.

Characteristic No. 1: Ideological Diversity. Ideologically diverse panels are more likely to produce accurate decisions than panels that are not ideologically diverse for at least two reasons. First, the judges on ideologically diverse panels are more likely to reach their decisions independently of each other, a key premise of the Condorcet Jury Theorem.[170] In contrast, panels that are not ideologically diverse are more likely to be susceptible to information cascades where judges follow one another instead of using independent judgment. [171] Accordingly, the votes on panels that are not ideologically diverse are more likely to be correlated, and the correlated votes for an inaccurate decision can

170. *See* Krishna K. Ladha, *The Condorcet Jury Theorem, Free Speech, and Correlated Votes*, 36 AM. J. POL. SCI. 617, 617 (1992).

171. *See id.*; David Austen-Smith & Jeffrey S. Banks, *Information Aggregation, Rationality, and the Condorcet Jury Theorem*, 90 AM. POL. SCI. REV. 34, 38–39 (1996); F. Andrew Hessick & Samuel P. Jordan, *Setting the Size of the Supreme Court*, 41 ARIZ. ST. L.J. 645, 695 (2009).

outnumber the uncorrelated votes for an accurate decision.[172] Second, ideologically diverse panels are more likely to adhere to precedent,[173] perhaps due to the presence of a potential dissenting judge who can "expose" the majority judges if they fail to do so.[174] The precedent-following decision is by definition the accurate decision when appellate panels, such as in the U.S. Courts of Appeals,[175] are duty bound to follow precedent.

Characteristic No. 2: Small in Size. While the Condorcet Jury Theorem suggests that larger judicial panels will be more accurate than smaller panels, and larger panels may have an advantage in that they can pool a greater amount of information, smaller panels have a key accuracy advantage over larger panels so long as Characteristic No. 1 is satisfied.[176] Judges on larger panels have more incentive to be "free-riders" because their individual votes are less likely to be decisive and there is more likely to be another judge who the free-riding judge perceives it can rely upon.[177] Accordingly,

---

172. *See* Hessick & Jordan, *supra* note 171, at 695 (arguing that diverse panels are more accurate because their votes are less likely to be correlated).

173. *See generally* Frank B. Cross & Emerson H. Tiller, *Judicial Partisanship and Obedience to Legal Doctrine: Whistleblowing on the Federal Courts of Appeals*, 107 YALE L.J. 2155 (1998) (finding that split partisan panels on the U.S. Courts of Appeals are more likely to follow precedent).

174. *See id.* at 2156 ("While there is undoubtedly more than one valid explanation for principled adherence to legal doctrine, we suggest that the prospect of a 'whistleblower' on the court—that is, the presence of a judge whose policy preferences differ from the majority's and who will expose the majority's manipulation or disregard of the applicable legal doctrine (if such manipulation or disregard were needed to reach the majority's preferred outcome)—is a significant determinant of whether judges will perform their designated role as principled legal decisionmakers.").

175. *See* Arthur D. Hellman, *Precedent, Predictability, and Federal Appellate Structure*, 60 U. PITT. L. REV. 1029, 1038 (1999) (noting that "all courts of appeals follow a rule under which panel decisions are binding on later panels unless overruled by the Supreme Court or by the court of appeals en banc").

176. *See generally* Albert B. Kao & Iain D. Couzin, *Decision Accuracy in Complex Environments Is Often Maximized by Small Group Sizes*, PROC. ROYAL SOC'Y, Mar. 18, 2014, at 1, 6 (arguing that small groups are able "to 'escape' the constraints of highly correlated information while retaining some of the benefits of pooling information collectively").

177. *See* Adrian Vermeule, *Many-Minds Arguments in Legal Theory*, 1 J. LEGAL ANALYSIS 1, 28 (2009) (arguing that larger groups may perform worse under the Condorcet Jury Theorem because the "free-riding problem is chronic

larger panels have the same problem as panels that are not ideologically diverse: information cascades where the inaccurate decision prevails because the correlated votes outnumber the uncorrelated votes. Finally, smaller panels are more likely to reach unanimous decisions. As further explained and argued below,[178] these unanimous decisions are more likely to be accurate because they are more likely to be reflective of the average view of the judges on the panel rather than the median view of the judges on the panel.

Characteristic No. 3: Majority Vote Combined With "Pressure" to Reach a Unanimous Decision. As noted above, appellate decisions generally have a binary component (which party wins) and a non-binary component (where on the spectrum the precedent is set). For the non-binary component, I contend that the most accurate placement on the spectrum is the placement that most represents the views of all the judges on the panel. Indeed, under one definition of accuracy in the legal context, a decision is considered accurate if it can be broadly considered as "better" than the alternatives.[179] It follows then that the placement that is *most* acceptable to the greatest number of judges can be construed as the *most* accurate. Phrased another way, the placement that can attract the most votes is the most "correct" placement because it is has the fewest number of judges saying it is "wrong."

Mathematically, the most representative placement of appellate decisions is typically going to be the "average" view of the judges on the panel because, unlike

---

and built into the very structure of the Theorem"). The free-riding problem is distinct from ordinary deliberation among judges, which likely does not violate the Condorcet Jury Theorem. *See Fitzpatrick*, *supra* note 49, at 111–12; Waldron, *supra* note 153, at 1715 n.76.

178. *See infra* Part III.B. (Characteristic No. 3).

179. *See* Hessick & Jordan, *supra* note 171, at 692; Fitzpatrick, s*upra* note 49, at 113 ("[F]or almost any question, we can imagine a range of answers that would be given in the federal judiciary; some of those answers will be more common and some of those answers will be less common. If the answer is less common, it is less accurate in the simple sense that most judges would have given a different answer.").

the median, it is a measure that incorporates the values of all the judges on the panel.[180] However, under a majority voting rule, one would expect that the view of the median judge—and not the average—would control the placement on the spectrum. Moreover, even if the voting rule required a supermajority or unanimity, this too is unlikely to lead to the placement on the spectrum of the average view.[181]

In order to achieve a spectrum placement of the average view in typical cases, the cases should be decided formally by majority vote but with informal pressure on the judges to reach a unanimous decision. While there should be significant incentives on the judges to reach a unanimous decision, the incentives should not be so overwhelming that they give the minority judges equal leverage with the majority judges on where the precedent is set on the spectrum. In other words, if the minority judges get too "greedy" in the negotiation, the majority judges should have the ability to go their own way and write their own opinion.[182] Under this type of framework, I contend that most decisions will "settle" with placement that correlates with the average view. And in cases where the judiciary is highly polarized—

180. *See generally* S. Manikandan, *Measures of Central Tendency: The Mean*, 2 J. PHARMACOLOGY & PHARMACOTHERAPEUTICS 140, 140 (2011) (noting that the "[c]entral tendency is defined as 'the statistical measure that identifies a single value as representative of an entire distribution'" and that the "[m]ean is the most commonly used measure of central tendency" as the "mean uses every value in the data and hence is a good representative of the data" (quoting FJ GRAVETTER & LB WALLNAU, STATISTICS FOR THE BEHAVIORAL SCIENCES (5th ed. 2000))). However, the mean is not the best measure of central tendency when there are "extreme values/outliers, especially when the sample size is small." *See id.* at 141.

181. A "rule" requiring a super-majority or unanimity may not be practical and to the extent it is practical will provide too much power to the minority judges who will, in theory, have equal power with the majority judges to shape the final decision. Therefore, the compromise point would be expected to be halfway between the majority and minority views, which is not the average view.

182. Having a mechanism that allows the majority judges to write separately solves the problem presented in atypical cases where the average view is not the most representative view of the panel, i.e., where one of the judges harbors a particularly extreme/outlier view. *See supra* note 180. In these atypical cases, the median view is more representative and (I argue) more accurate.

e.g., there is a significant gap between the most "liberal" Republican-appointed judge and the most "conservative" Democrat-appointed judge—one would expect the "average" view to be more moderate than the "median" view. For example, imagine rating judges on an ideological score of ten to one, with a higher score denoting a more conservative judge, a lower score denoting a more liberal judge, and a score of 5.5 denoting a perfectly ideologically centrist judge. If the judges on a three-judge panel have ideological scores of eight, seven, and three, the view of the median judge would be a seven but the "average" view of the judges on the panel would be a six.[183]

We have real-world data indicating that these three characteristics are present at the U.S. Courts of Appeals when an ideologically diverse three-judge panel is formed via random selection. When this happens, the panel almost always reaches a unanimous decision,[184] even though the empirical data is clear that the individual judges do, in fact, hold differing views on these cases, even where the case does not appear to be politically charged.[185] The panel feels pressure to reach a unanimous decision because, among other reasons,[186]

183. Where there is less polarization, one would expect the median and mean to be closer.

184. Lee Epstein et al., *Why (and When) Judges Dissent: A Theoretical and Empirical Analysis*, 3 J. LEGAL ANALYSIS 101, 106, 106 n.9 (2011) (finding a dissent rate at the U.S. Courts of Appeals of just 2.6% during the 1990 to 2007 time frame).

185. *See* Alma Cohen, *The Pervasive Influence of Ideology at the Federal Circuit Courts* (Nat'l Bureau of Econ. Rsch., Working Paper No. 31509, 2023) (analyzing a dataset of about 600,000 cases from the U.S. Courts of Appeals from 1985 to 2020 and finding that "the political affiliation of judges is associated with outcomes, and thus can help to predict them, throughout the vast universe of [such] cases—and not only in the ideologically contested cases"); *see also* LEE EPSTEIN ET AL., THE BEHAVIOR OF FEDERAL JUDGES: A THEORETICAL AND EMPIRICAL STUDY OF RATIONAL CHOICE 7 (2013); CASS R. SUNSTEIN ET AL., ARE JUDGES POLITICAL? AN EMPIRICAL ANALYSIS OF THE FEDERAL JUDICIARY viii (2006); Richard L. Revesz, *Environmental Regulation, Ideology, and the DC Circuit*, 83 VA. L. REV. 1717, 1717–19 (1997).

186. Of course, smaller panels are also more likely to reach unanimous decisions because there are fewer judges that might disagree. *See* Epstein et al., *supra* note 184, at 108.

the workload costs of issuing a dissenting opinion[187]—which falls on not just the dissenting judge but also on the majority judges who now need to respond to the minority judge's arguments[188]—are high. Accordingly, the minority judge has leverage to shape the final decision.

Indeed, there are empirical studies showing that decisions across many areas of law are more moderate if there are two Republican appointees (R) or two Democratic appointees (D) on three-judge panels rather than three judges of the same party.[189] If all that mattered was the view of the median judge, the decisions of RRD panels should not be significantly less conservative than the decisions of RRR panels and the decisions of RDD panels should not be significantly less liberal than the decisions of DDD panels, which suggests that the minority judge is able to move the majority judges off the "median" position to a position more reflective of the "average."[190]

---

187. This is referred to in the literature as the "dissent aversion" theory. *See* RICHARD A. POSNER, HOW JUDGES THINK 31–34 (2008).

188. *See* Epstein et al., *supra* note 184, at 102 ("A dissent in the court of appeals increases the length of the majority opinion by about 20 percent, which we treat as a rough measure of the cost that a dissent imposes on the majority.").

189. *See, e.g.*, SUNSTEIN ET AL., *supra* note 185, at 12 (finding "a large disciplining effect . . . from the presence of a single panelist from another party" on three-judge panels at the U.S. Courts of Appeals as "all-Republican panels show far more conservative patterns than majority Republican panels, and all-Democratic panels show far more liberal patterns than majority Democratic panels"); EPSTEIN ET AL., *supra* note 185, at 191 ("Confirming the earlier empirical literature, we find that the presence on a panel of a judge appointed by a President of a different party from that of the President (or Presidents) who appointed the other judges on the panel tends to moderate the voting of those judges."); *see also* Sean Farhang, *Supreme Court Oversight of the Federal Rules: A Principal-Agent Problem?*, 72 DEPAUL L. REV. 363, 381 (2022) (reviewing the empirical literature and concluding that "when court of appeals judges' party, gender, and race are associated with votes, judges in the preference-minority on panels regularly influence the outcome votes of judges in the preference-majority").

190. To be sure, the minority judge will not always be able—or even desire—to move the ultimate decision from the "median" position to the "average" position. For example, if the majority judges care a lot about the particular case, they may be unwilling to compromise, and if the minority judge believes that a higher court is likely to reverse the panel decision, the minority judge may opt to write a dissent rather than make the majority opinion "less bad." Indeed, in

However, the pressure to reach a unanimous decision is greatest where the panels are small and the caseload is heavy. In larger panels, the pressure is much less because the workload costs of issuing a dissenting opinion can be absorbed by more judges and larger panels generally hear fewer cases.[191] The spectrum placement for larger panels is likely therefore to be the view of the median judge—or even the median judge of the majority coalition.[192]

Accordingly, I contend that the accuracy of the decisions of the U.S. Supreme Court could be improved if, for example, the Court, currently consisting of six Republican appointees and three Democratic appointees, made its decisions in panels of three, with each panel consisting of two randomly selected Republican appointees and one randomly selected Democratic appointee, particularly if the number of cases that the Court heard significantly increased without a

---

certain situations, if the would-be minority judge feels very strongly about the case and the case is not a priority for the would-be majority judges, the final decision might be closer to the minority judge's position than the majority judges' position. *See* Epstein et al., *supra* note 184, at 108 ("If one judge feels strongly that the case should be decided one way rather than another, while the other two judges, though inclined to vote the other way, do not feel strongly, one of those two may decide to go along with the third to avoid creating ill will, perhaps hoping for reciprocal consideration in some future case in which he has a strong feeling and the other judges do not."); *see also* Jonathan P. Kastellec, *Racial Diversity and Judicial Influence on Appellate Courts*, 57 AM. J. POL. SCI. 167, 179 (2013) (finding that "the random assignment of a black judge to a three-judge panel in affirmative action cases nearly ensures that the panel will issue a liberal decision").

191. For example, there were dissents in 62% of the cases at the U.S. Supreme Court, compared to just 2.6% of the cases at the U.S. Courts of Appeals, during the time frame of 1990 to 2007. *See* Epstein et al., *supra* note 184, at 106 & n.9.

192. While social scientists have long thought that the outcome of the U.S. Supreme Court cases is based on the view of the median justice, *see* Andrew D. Martin et al., *The Median Justice on the United States Supreme Court*, 83 N.C. L. REV. 1275, 1278 (2005), more recent research suggests that U.S. Supreme Court decisions are most likely to reflect the view of the median justice in the majority coalition, which would make the decision even less likely to represent the average view of the justices. *See* Tom S. Clark & Benjamin Lauderdale, *Locating Supreme Court Opinions in Doctrine Space*, 54 AM. J. POL. SCI. 871, 872 (2010) (finding evidence that the view of the "median justice in the majority coalition most powerfully predicts majority opinion location" for U.S. Supreme Court decisions).

commensurate increase in its support staff.[193] The goal of this structural reform would be less polarized 6-3 or 5-4 decisions. Instead, these polarized decisions would be replaced by more moderate unanimous three-judge opinions, shaded to the majority side of the ideological spectrum but not overwhelmingly so, and more reflective of the view of the average Justice. Notably, some of the Justices have themselves suggested that such "consensus" opinions have added authority not found in split decisions.[194]

Based on this premise—and I realize that I am entering the realm of science fiction for the moment—an AI U.S. Supreme Court could be created with nine robot justices that replicate the views of the nine current Justices. This AI Court would hear cases in panels of three ideologically diverse robot justices, reflective of the ideological makeup of the human U.S. Supreme Court, and under a majority voting decision rule. The robot justices would be programmed to deliberate, and negotiate, over the case outcome. As it is not possible to "pressure" the robot justices to reach unanimous decisions due to workload costs, the robot justices would need to be programmed in a way that incentivizes them

193. This proposal is a combination of two previous proposals in the literature: George and Guthrie's proposal that the U.S. Supreme Court decides its cases in panels with en banc review, *see* Tracey E. George & Chris Guthrie, *Remaking the United States Supreme Court in the Courts' of Appeals Image*, 58 DUKE L.J. 1439, 1442 (2009), and Tiller and Cross's proposal that split partisan panels be mandated at the U.S. Courts of Appeals, *see* Emerson H. Tiller & Frank B. Cross, *A Modest Proposal for Improving American Justice*, 99 COLUM. L. REV. 215, 215 (1999). To account for the fact that not every Democrat-appointed Justice and Republican-appointed Justice will follow the party line on every issue, the Justices could also be assigned numerical scores based on where they fall on the ideological spectrum for particular issues. *See* Michael J. Hasday, *The Rank-Order Method for Appellate Subset Selection*, 93 NOTRE DAME L. REV. ONLINE 17, 22 (2017).

194. *See* Jeffrey Rosen, *Roberts's Rules*, ATLANTIC (Jan. 1, 2007, 12:00 PM), http:// www.theatlantic.com/magazine/archive/2007/01/robertss-rules/305559/. For example, in an interview, Chief Justice Roberts contended that "[u]nanimous, or nearly unanimous, decisions are hard to overturn[,] contribute to the stability of the law" and garner more public respect. *Id.*; *see also* Trump v. Anderson, 601 U.S. 100, 118 (2024) (Barrett, J., concurring) (stressing that "the message Americans should take home" is that "[a]ll nine Justices agree[d] on the outcome of this case").

to reach unanimous decisions. For example, each robot justice could be programmed to play a "game" where they seek to collect points for each assigned case, with points awarded based on how close the controlling opinion is to that robot justice's view, how important the case is to that robot justice, if the robot justice is part of a panel that decides the case unanimously, and perhaps other factors. The precise specifications of the algorithm—e.g., exactly how many points should be awarded for unanimous decisions—would need to be determined, but the degree of pressure faced by a U.S. Court of Appeals judge in three-judge panels could be used as a starting point. In short, this proposal aims to create a "virtual" environment to aggregate the preferences of the Justices on the U.S. Supreme Court because the real-world environment is sub-optimal.

The human U.S. Supreme Court would still need to hear cases to provide up-to-date data for their robot justice counterparts. There are at least two different ways for this to be accomplished. Every decision by the panels of three ideologically diverse robot justices could be subject to en banc review by the human Justices, similar to the structure of the U.S. Courts of Appeals. In addition, or alternatively, based on the subject matter of the case—or perhaps even randomly—certain cases could be decided by panels of three ideologically diverse robot justices while other cases could be decided by panels of three ideologically diverse human Justices or by the nine human Justices en banc.

While this proposal is probably not technologically feasible today, researchers are actively exploring the ability of AI to engage in "negotiation games" and finding early indications that it is able to do so.[195] Indeed, AI

195. *See, e.g.*, Yao Fu et al., Improving Language Model Negotiation with Self-Play and In-Context Learning from AI Feedback, at 1, 9 (May 17, 2023), ARXIV:2305.10142, https://arxiv.org/pdf/2305.10142 [https://perma.cc/CA5Y-MTX4] (studying "whether multiple large language models can autonomously improve each other in a negotiation game by role-playing and learning from AI feedback" and finding "that certain models can indeed improve by continuously playing competition games with iterative AI feedback, under well-defined rules").

robots are currently able to defeat professional poker players in poker,[196] a game that has been likened to the "game" that judges play in multi-judge, appellate panels.[197]

IV. CONCLUSION

Much like Chief Justice Roberts, I expect that there are many people who are skeptical of the concept of robot judges and who view the role of robots, if any, as assisting human judges and not replacing them.[198] Even if that is your view, I contend that it is still worth contemplating how robot judges can be shown to be more accurate than human judges as it might improve the accuracy of (human) legal decision making without the robots deciding actual cases.

Indeed, somewhat lost in the debate about robot umpires is that since the development of the technology that makes robot umpires possible, *human* umpires have

196. *See* Bernard Marr, *Artificial Intelligence Masters the Game of Poker—What Does That Mean for Humans?*, FORBES (Sept. 13, 2019, 12:43 AM), https://www.forbes.com/sites/bernardmarr/2019/09/13/artificial-intelligence-masters-the-game-of-poker—what-does-that-mean-for-humans/?sh=10d85b2a5f9e [https://perma.cc/BG7Q-JPM8] ("While AI had some success at beating humans at other games such as chess and Go (games that follow predefined rules and aren't random), winning at poker proved to be more challenging because it requires strategy, intuition, and reasoning based on hidden information. Despite the challenges, artificial intelligence can now play—and win—poker.").

197. *See* Diane P. Wood, *When to Hold, When to Fold, and When to Reshuffle: The Art of Decisionmaking on a Multi-Member Court*, 100 CALIF. L. REV. 1445, 1447 (2012). Indeed, one U.S. Court of Appeals judge wrote an article making this comparison and said:

> What should a judge do when she disagrees with her colleagues on the bench? Should the potential dissenter always "fold"? Should the potential dissenter adamantly refuse to meet others halfway? There is no singular answer to these questions. Rather, the answer depends on criteria like a judge's aversion to the extra work that a separate opinion entails, her desire to get along with her colleagues, or the purity of her intentions.

Id. at 1475.

198. *See* ROBERTS, *supra* note 8, at 6 ("I predict that human judges will be around for a while.").

become much more accurate.[199] It appears, at least in the professional baseball context, that just having the tools to measure accuracy—and perhaps the threat of competition—can by itself improve accuracy. There may be a similar effect on human judges in the traditional legal setting, which raises some intriguing possibilities.

For example, in the comparison path discussed in Part II, the U.S. legal system could be more proactive in measuring and publicizing the accuracy rate of human judges using the "reversal rate" as the metric. Moreover, if technology develops that allows robot judges to be better at this metric than human judges (or at least better than the median human judge), one could imagine a scenario in which human judges are given the robot judge's predictions about how certain decisions will fare at the next level of the judicial hierarchy, allowing the human judges to identify the straightforward cases and focus their attention on cases in which the robot judge has less predictive certainty.[200] It also may be possible to analyze whether human judges *assisted* by robot judges can achieve a better reversal rate than either human judges or robot judges acting on their own.

High courts present a particularly vexing problem for measuring accuracy because the "reversal rate" metric is of course not available. However, the decision making of high courts might be perceived as more accurate in the aggregate if most of their decisions are unanimous or near unanimous and less accurate in the aggregate if too many of their decisions are divided and several of the judges are in effect saying they are wrongly decided. The problem is that the human judges might not be properly incentivized to form a consensus. Here, the

199. *See* Andrews, *supra* note 24 ("Since the beginning of the pitch tracking era in 2008, [human] umpires have improved their accuracy in calling balls and strikes every single year. Accuracy has gone from 81.3% to 92.4%. If an improvement of 11.1% in 15 years doesn't sound particularly big, consider it this way: incorrect calls have been cut by nearly 60%.").

200. In a similar vein, Ryan Copus has argued that the U.S. Courts of Appeals should use machine learning techniques to predict the probability of reversal for lower court decisions in order to identify, and allocate judicial attention on, the hard cases. *See* Copus, *supra* note 131, at 605–06.

AI Supreme Court proposed in Part III might be used to test various structural reforms of the human Supreme Court to achieve a more optimal compromising culture. The AI Supreme Court could also be employed to show the human Supreme Court avenues to compromise when the human Justices are "stuck" in their ideological camps.[201]

In short, the various paths discussed in this article may have applications outside the human vs. robot framework. They may also light the path for how *human* judges can be better.

201. *See* Unikowsky, *supra* note 130.