Having Your Day in Robot Court

Chen, MB
Stremitzer, A
Tobia, K

University of Hong Kong Faculty of Law Research Paper No. 2021/020

HAVING YOUR DAY IN ROBOT COURT

Benjamin Minhao Chen,* Alexander Stremitzer,** and Kevin Tobia***

Abstract: Should machines be judges? Some say no, arguing that citizens would see robot-led legal proceedings as procedurally unfair because "having your day in court" is having another human adjudicate your claims. Prior research established that people obey the law in part because they see it as procedurally just. The introduction of artificially intelligent (AI) judges could therefore undermine sentiments of justice and legal compliance if citizens intuitively take machine-adjudicated proceedings to be less fair than the human-adjudicated status quo. Two original experiments show that ordinary people share this intuition. There is a perceived "human-AI fairness gap."

However, it is also possible to reduce — and perhaps even eliminate — the fairness gap through "algorithmic offsetting." Affording a hearing before AI judges and enhancing the interpretability of AI-rendered decisions reduce the human-AI fairness gap. Moreover, the procedural justice advantage of a human over AI appears to be driven more by beliefs about the accuracy of the outcome and thoroughness of consideration, rather than doubts about whether a party felt it had a good opportunity to voice its opinions or whether the judge understood the perspective of the litigant.

The results support a common and fundamental objection to robot judges: There is a concerning human-AI fairness gap. Yet, the results also indicate that the strongest version of this challenge — human judges have irreducible procedural fairness advantages — is not reflected in public views. In some circumstances, people see a day in robot court as no less fair than a day in human court.

* Assistant Professor of Law, University of Hong Kong.
** Professor of Law, Economics, and Business, ETH Zurich and Visiting Professor of Law, UCLA Law School.
*** Associate Professor of Law, Georgetown University Law Center.
We are grateful to Henry Kim for statistical advice; to Elliott Ash, Dan Ho, Aileen Nielsen, and Richard Re for helpful comments; and Noemi Birchler, Jean Chang, Nils Heinemann, Costanza Maria Improta, Marine Jorio, and Jannek Ulm for excellent research assistance.

Table of Contents

INTRODUCTION

"Can you foresee a day when smart machines, driven with artificial intelligences, will assist with courtroom factfinding or, more controversially even, judicial decision-making?"[1] Shirley Ann Jackson, a college president and theoretical physicist, posed this question to Chief Justice John Roberts in 2017. The Chief Justice's answer: "It's a day that's here."[2]

Artificial intelligence (AI) has now found applications in the U.S. legal system. Thus far, AI has primarily served as an aid, not an ultimate decisionmaker. For example, algorithms recommend but do not determine criminal sentences in some states.[3] Nevertheless, AI could function as primary decisionmakers in some administrative contexts, such as terminating welfare benefits or targeting people for air travel exclusions.[4] Outside the United States, there are plans to give greater judicial decision-making responsibility to machines.[5] Estonia is piloting AI adjudication of some small claims.[6] China has declared the integration of AI into judicial processes a national priority, introducing, for example, precedent recommendation systems that assist human judges by formulating judgments based on past decisions.[7]

As technological advances make robot judging a possibility, new legal-ethical challenges arise. Perhaps the most critical objection sounds in procedural fairness. Would a judicial proceeding overseen by a robot judge

---

[1] Adam Liptak, *Sent to Prison by a Software Program's Secret Algorithms*, N.Y. TIMES (May 1, 2017), https://www.nytimes.com/2017/05/01/us/politics/sent-to-prison-by-a-software-programs-secret-algorithms.html [https://perma.cc/3TG6-62FH].

[2] *Id.*

[3] *See id.*; *cf.* Frank Fagan & Saul Levmore, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CALIF. L. REV. 1, 1 (2019) (arguing that AI's role in criminal, corporate, and contract law rules has empirical limitations). Of course, AI is also increasingly the object of law. *See, e.g.*, Jeffrey J. Rachlinski & Andrew J. Wistrich, *Judging Autonomous Vehicles*, YALE J.L. & TECH. (forthcoming), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3806580 [https://perma.cc/22G9-HTZU].

[4] Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1252 (2008); *see also* Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147, 1213–21 (2017) (discussing how AI could aid in the efficient administration of the state).

[5] Although the words robot and machine can be taken as referring to a physical device as opposed to a sequence of rules or operations for deriving outputs from inputs this Article uses the terms artificial intelligence, algorithm, machine, and robot interchangeably when describing adjudication by non-human, computational systems.

[6] Eric Niiler, *Can AI Be a Fair Judge in Court? Estonia Thinks So*, WIRED (Mar. 25, 2019, 7:00 AM), https://www.wired.com/story/can-ai-be-fair-judge-court-estonia-thinks-so [https://perma.cc/2PVW-PA33].

[7] *See* Ray Worthy Campbell, *Artificial Intelligence in the Courtroom: The Delivery of Justice in the Age of Machine Learning*, 18 COLO. TECH. L.J. 323, 343 (2020); Jinting Deng, *Should the Common Law System Welcome Artificial Intelligence: A Case Study of China's Same-Type Case Reference System,* 3 GEO. L. TECH. REV. 223, 224–26 (2019).

undermine the right to a fair trial?[8] This concern can be articulated doctrinally: Does robot judging violate the European Convention on Human Rights' fair trial standards or constitutional commitments to due process?[9] The concern can also be articulated in legal-ethical terms. Assuming the doctrinal hurdles are overcome, would people reject robot judging as procedurally unfair?

This Article enters the debate from this second perspective, considering people's judgments of procedural fairness. A long tradition in legal psychology has studied the social psychology of procedural justice in this way.[10] Evidence suggests that the perceived fairness of legal processes has important practical implications. People obey the law, in part, because it is seen to be fair.[11] The public's assessment of the fairness of robot judges is therefore crucial, both for those concerned with legal compliance and for those who ascribe intrinsic value to ordinary citizens' conceptions and experiences of fairness.

Fairness and procedural legitimacy are at the heart of modern debates about AI judging. As Campbell puts it, "[i]n asking whether AI can play the role of judges, we must ask . . . . [whether] AI courts can enable public participation, give participants a sense of being fairly heard . . . [and] vindicate the legitimacy not just of the courts, but of the governmental systems within which they reside."[12]

Re and Solow-Niederman articulate a similar concern, noting that "the incomprehensibility of an AI adjudicator could pose legitimacy or fairness problems for individuals who are subjects of AI adjudication . . . . The individual without comprehension might thus experience special or separate [procedural] harms." [13] Even in discussions about alternative dispute resolution, perceived procedural fairness matters. For example, a central criterion in assessing whether computers can "be fair" in online dispute resolution is "disputants' evaluation of the fairness of . . . [the] process."[14]

Whether people see robot judges as fair is a largely unexplored empirical question.[15] Through a series of original experimental studies involving a large

---

[8] *See*, *e.g.*, Aleš Završnik, *Criminal Justice, Artificial Intelligence Systems, and Human Rights*, 20 J. ACAD. EUR. L. 567, 576–78 (2020); *see generally* Maria Dymitruk, *The Right to a Fair Trial in Automated Civil Proceedings*, 13 MASARYK U. J.L. & TECH. 27 (2019).

[9] *See* Loomis v. Wisconsin, 881 N.W.2d 749, 760–64 (Wis. 2016) (holding that a trial court's use of an algorithmic risk assessment does not violate due process rights), *cert. denied*, 137 S. Ct. 2290 (2017); Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 621–26 (2020) (cataloguing constitutional and other legal impediments to machine judgment).

[10] *See* E. ALLAN LIND & TOM R. TYLER, THE SOCIAL PSYCHOLOGY OF PROCEDURAL JUSTICE 11–15 (1988).

[11] *See* TOM R. TYLER, WHY PEOPLE OBEY THE LAW 19–29 (2006).

[12] Campbell, *supra* note 7, at 341.

[13] Richard M. Re & Alicia Solow-Niederman, *Developing Artificially Intelligent Justice*, 22 STAN. TECH. L. REV. 242, 264 (2019).

[14] Ayelet Sela, *Can Computers Be Fair?*, 33 OHIO ST. J. ON DISP. RESOL. 91, 105 (2018).

[15] There are extant studies of blameworthiness and responsibility judgment about scenarios involving AI, or AI and humans. *See, e.g.*, Edmond Awad et al., *Drivers Are Blamed More Than Their Automated Cars When Both Make Mistakes*, 4 NATURE HUM.

sample of U.S. participants, we present evidence of people's evaluation of robot judges. The experiments vary the decisionmaker (human or AI), scenario (bail, sentencing, or consumer arbitration), whether there is a hearing, and whether the judge's decision is interpretable.[16]

The study makes two important findings. First, there is a clear "human-AI fairness gap": proceedings conducted by human judges are seen as fairer than their AI counterparts. Second, the procedural fairness advantage of human judges is, in all likelihood, neither irreducible nor absolute. In fact, our results hint at the possibility of "algorithmic offsetting." That is, the human-AI fairness gap can be offset, partly and potentially entirely, by introducing into machine adjudication procedure that might be absent from actual, human adjudication, elements such as a hearing or an interpretable decision. Surprisingly, participants did not evaluate a hearing before an AI judge as meaningless. To the contrary, having the opportunity to speak and be heard increases procedurally fairness ratings for both human and AI-adjudicated processes.

Moreover, an exploratory mediation analysis suggests that the human-AI fairness gap is explained more by "hard" factors like the perceived accuracy and thoroughness of the decision-making process, rather than distinctively human, "soft" factors, like the decision-maker's understanding of the defendant's position or a feeling that the defendant had a voice. This finding suggests that in domains where quantitative information about a decision's accuracy is available, the superior accuracy of algorithms may eventually erode or even eliminate the fairness gap.

The final part of the Article develops implications from these findings. We elaborate on the idea of algorithmic offsetting: Closing the human-AI fairness

---

BEHAV. 134 (2020); Gabriel Lima et al., *Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision Making*, PROC. 2021 CONF. ON HUM. FACTORS COMPUTING SYS. 1, 1 (2021); Bertram Malle et al., *Sacrifice One for the Good of the Many? People Apply Different Moral Norms to Human and Robot Agents*, 10 ACM/IEEE INT'L CONF. ON HUM.-ROBOT INTERACTION 117, 117 (2015); Gabriel Lima & Meeyoung Cha, *Human Perceptions of AI-Caused Harm*, The CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURIS. (forthcoming 2024) (on file with authors). Other scholars have studied judgments about legal standards related to AI tools in other contexts. *See, e.g.*, Kevin Tobia, Aileen Nielsen, & Alexander Stremitzer, *When Does Physician Use of AI Increase Liability*, 62 J. NUCLEAR MED. 17, 17 (2021).

[16] Interpretability, roughly speaking, refers to "the ability to explain or to present in understandable terms to a human," Finale Doshi-Velez & Been Kim, Towards a Rigorous Science of Interpretable Machine Learning 2 (Mar. 2, 2017) (unpublished manuscript) (on file with authors). Some authors treat interpretability and explainability as synonyms. *See* Ricards Marcinkevics & Julia E. Vogt, Interpretability and Explainability: A Machine Learning Zoo Mini-tour 1 (Dec, 2020) (unpublished manuscript) (on file with authors). Others distinguish between interpretable machine learning—where the models use are "inherently interpretable"—and explainable machine learning where "post-hoc" models are developed to explain functions "that [are either] too complicated for any human to comprehend or . . . [are] proprietary." Cynthia Rudin, *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, 1 NATURE MACH. INTELLIGENCE 206, 206 (2019). We follow this distinction here.

gap by issuing AI decisions that are more interpretable than human-rendered decisions, or by offering litigants a meaningful hearing before an AI judge when they would not have had such an opportunity in a human-adjudicated proceeding. The empirical results suggest that people evaluate AI judging under such circumstances as being as procedurally fair as human judging. And, as Eugene Volokh puts it, "[o]ur question should not be whether AI judges are perfectly fair, only whether they are at least as fair as human judges."[17]

It might seem that "having your day in court" requires being heard before a human judge and anything else violates fairness. Insofar as human judges set the yardstick for fairness, our results indicate that the procedural justice objection against AI robot judges may not be a decisive one. In fact, our results suggest that were robot judges to become more accurate, comprehensive, interpretable, or responsive, their decision-making might even be seen as fairer than that of some human judges.

I. AUTOMATING THE JUDICIARY AND PROCEDURAL LEGITIMACY

Should machines decide cases? While commentators describe the rise of AI in epochal terms, the thought that robots might one day settle legal disputes is hardly new. In 1977, the human rights scholar Anthony D'Amato mused that computers might replace judges, assuming that "the law has been made completely determinable" and automation eliminates discretion in judicial decision-making.[18] Law has not become completely determinable. Nor is it likely to. Legal language is open-textured[19] and the rivalry between textualism, intentionalism, and purposivism persists when it comes to statutory interpretation.[20] Meanwhile, the evaluative nature of many common law concepts means that the application of old wisdom to new problems remains an exercise in normative reasoning. Instead of repudiating human judgment, state-of-the-art computers strive to replicate it.[21] Leveraging greater computing power and more flexible modelling strategies, modern algorithms identify and harness empirical relationships more effectively than their predecessors.[22]

---

[17] Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1169 (2019).

[18] Anthony D'Amato, *Can/Should Computers Replace Judges*, 11 GA. L. REV. 1277, 1279 (1977).

[19] Frederick Schauer, *On the Open Texture of Law*, 87 GRAZER PHILOSOPHISCHE STUDIEN 197, 202 (2013).

[20] *See, e.g.*, Frank H. Easterbrook, *The Absence of Method in Statutory Interpretation*, 84 U. CHI. L. REV. 81, 81–82 (2017); Kevin P. Tobia, *Testing Ordinary Meaning*, 134 HARV. L. REV. 726, 728–29 (2020). *See generally* Elias Leake Quinn, *What if Big Data Helped Judges Decide Exactly What Words Mean?*, SLATE (Apr. 8, 2021, 2:00 PM), https://slate.com/technology/2021/04/corpus-linguistics-algorithmic-bias-judicial-opinions.html [https://perma.cc/ZY6T-XEFF] (suggesting that algorithms will not resolve all legal interpretive questions).

[21] *See* Edmond Awad et al., *Computational Ethics*, 26 TRENDS COGNITIVE SCIS. 388, 392 (2022).

[22] *See generally* Stuart Nagel, *Predicting Court Cases Quantitatively*, 63 MICH. L. REV. 1411 (1965).

Simple models have already outperformed lawyers in predicting decisions of the United States Supreme Court,[23] and more sophisticated models are now boasting impressive accuracy for a diverse range of tribunals.[24] Their apparent success has excited interest in the possibility of faster, cheaper, and better justice delivered by machines rather than humans.

The role of AI in American criminal law remains very much advisory — legal judgment continues to be delivered by judges sitting in courtrooms. But in the United Kingdom, the public law barrister Lord Pannick has wondered "whether consistency in sentencing decisions might be promoted, irrelevant factors excluded, and a lot of money saved on sentencing appeals by the use of a computer programme."[25] And while no jurisdiction has to date been bold enough to let computers alone determine a person's guilt or innocence, at least one nation is prepared to let some kinds of cases be resolved by machines. Estonia is building a robot to adjudicate small claims where the amounts in controversy are below €7,000.[26] According to the chief data scientist on the project, Ott Velsberg, the country is hospitable ground for such an experiment given that its 1.3 million residents are accustomed to the digitization of public services like voting and tax filing.[27]

These developments raise questions about human adjudication's distinctiveness — and questions about its future. From a theoretical perspective, adjudication has never been solely about achieving the right result. Lon Fuller, for example, characterized adjudication as a form of social ordering distinguished by "the fact that it confers on the affected party a peculiar form of participation in the decision, that of presenting proofs and reasoned arguments for a decision in his favor."[28] "Whatever heightens the significance of this participation lifts adjudication towards its optimum expression" and "[w]hatever destroys the meaning of that participation destroys the integrity

---

[23] Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin & Kevin M. Quinn, *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decision-Making*, 104 COLUM. L. REV. 1150, 1150 (2004).

[24] *See* Daniel Martin Katz et al., *A General Approach for Predicting the Behavior of the Supreme Court of the United States*, PLOS ONE (Apr. 2017), https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0174698&type=printable [https://perma.cc/K4X6-W27U]; Nikolaos Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, PEERJ COMPUT. SCI. (Oct. 24, 2016), https://peerj.com/articles/cs-93 [https://perma.cc/ZRQ9-RYQA]; Masha Medvedeva, Michel Vols & Martijn Wieling, *Using Machine Learning to Predict Decisions of the European Court of Human Rights*, 28 A.I. & L. 237, 237 (2019); Andre Lage-Freitas et al., Predicting Brazilian Court Decisions (Apr. 20, 2019) (unpublished manuscript) (on file with author).

[25] David Pannick, *Why No Offender Wants to Face a Judge Who is Tired, Hungry, or Disappointed*, THE TIMES (Jan. 19, 2017), https://www.thetimes.co.uk/edition/law/why-no-offender-wants-to-face-a-judge-who-is-tired-hungry-or-disappointed-6bdxbm2w0 [https://perma.cc/ERK5-YNKR].

[26] Niiler, *supra* note 6.

[27] David Cowan, *Estonia: A Robotically Transformative Nation*, ROBOTICS L.J. (July 26, 2019), https://www.roboticslawjournal.com/global/estonia-a-robotically-transformative-nation-28728942 [https://perma.cc/MB78-Y7DC].

[28] Lon Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353, 364 (1978).

of adjudication itself."[29] To the extent, then, that machines are unable to respond to reason, automated adjudication is an oxymoron.

Whether or not Fuller is correct about the essence of adjudication, the procedural dimension to the rule of law calls for subjects to be accorded the opportunity to interpret the law, relate its abstract demands to their own circumstances, and have their arguments evaluated impartially in a neutral forum.[30] These procedural guarantees, Jeremy Waldron argues, are at the heart of ordinary understandings of legality.[31] "They capture a deep and important sense associated foundationally with the idea of a legal system, that law is a mode of governing people that treats them with respect, as though they had a view or perspective of their own to present on the application of the norm to their conduct and situation."[32] On this view, the advent of robot judges who compute but do not contemplate threatens to undermine the rule of law as it is popularly conceived.

Psychological research has documented the importance of procedure for people's experiences and perceptions of fairness.[33] While early studies addressed the consequences of unequal resource allocations on attitudes and behavior, later contributions examined how those allocations were made, concluding that form is sometimes as critical as substance.[34] The shift in emphasis from distributive to procedural justice brought about an accompanying change in paradigm — from one focused on outcomes to one centered on relationships.[35] Procedures are valued because they allow parties to convey information to the adjudicator.[36] Procedures are also valued because they treat the parties not as objects but as subjects who have an interest to defend and a perspective to offer.[37] Litigants who believe they have received

---

[29] *Id.*

[30] *See* Jeremy Waldron, *The Concept and the Rule of Law*, 43 GA. L. REV. 1, 6–9 (2008) (describing the normative and procedural aspects of the rule of law).

[31] *See* Jeremy Waldron, *The Rule of Law and the Importance of Procedure*, *in* GETTING TO THE RULE OF LAW: NOMOS L 3, 3–16 (James E. Fleming ed., 2011).

[32] *Id.* at 15–16.

[33] Tom R. Tyler & E. Allan Lind, *Procedural Justice*, *in* HANDBOOK OF JUSTICE RESEARCH IN LAW 65, 66–68 (Joseph Sanders & V. Lee Hamilton, eds., 2001). To be clear, this approach is descriptive-explanatory, not normative-prescriptive. Researchers in this tradition investigate how ordinary people experience justice and fairness; they do not pass judgment on the truth of lay people's understandings. Gerold Mikula, *Some Observations and Critical Thoughts About the Present State of Justice Theory and Research*, *in* WHAT MOTIVATES FAIRNESS IN ORGANIZATIONS? 197, 198–99 (Stephen W. Gilliland et al. eds., 2005).

[34] E. Allan Lind, *The Study of Justice in Social Psychology and Related Fields*, *in* SOCIAL PSYCHOLOGY AND JUSTICE 1, 6 (E. Allan Lind ed., 2019).

[35] *See* Tom R. Tyler, *Social Justice: Outcome and Procedure*, 35 INT'L J. PSYCH. 117, 118–20 (2000).

[36] John Thibault & Laurens Walker, *A Theory of Procedure*, 66 CALIF. L. REV. 541, 551 (1978).

[37] Tom R. Tyler, *Psychological Models of the Justice Motive: Antecedents of Distributive and Procedural Justice*, 67 J. PERSONALITY & SOC. PSYCH. 850, 852–52 (1994); Tom R. Tyler & Steven L. Blader, *The Group Engagement Model: Procedural Justice, Social Identity, and Cooperative Behavior*, 7 PERSONALITY & SOC. PSYCH. REV. 349, 351 (2003) (finding quality of treatment to be a key input in judgments of procedural fairness).

procedural justice are more likely to recognize the authority of the tribunal and accept its determination.[38] While there are several factors conducive to a sense of fairness,[39] two of them are especially relevant to the case of AI judgments: voice and justification.

First, people are more inclined to endorse a procedure as fair if they are able to voice their opinions.[40] Voice matters for instrumental and value-expressive reasons. Instrumentally, the chance to suggest or advocate a position gives the speaker possible influence over outcomes.[41] Hence, those who have voice may regard a process as fair because their views could shape the decisions being made.[42] But they may also regard a process as fair even when their opinions have little hold on the decisionmaker.[43] This is because the opportunity to speak acknowledges the parties' agency and their membership in the community.[44] The denial of such an opportunity is especially aggravating in societies and situations where it is expected,[45] and

---

[38] *Cf.* Tom R. Tyler & Kenneth Rasinski, *Procedural Justice, Institutional Legitimacy, and the Acceptance of Unpopular U.S. Supreme Court Decisions: A Reply to Gibson*, 25 LAW & SOC'Y REV. 621 (1991) (finding that the public's views about the fairness of U.S. Supreme Court procedures influences its view of the Court's authority); *see also* Stanislaw Burdziej, Keith Guzik & Bartosz Pilitowski, *Fairness at Trial: The Impact of Procedural Justice and Other Experiential Factors on Criminal Defendants' Perceptions of Court Legitimacy in Poland*, 44 LAW & SOC. INQUIRY 359 (2019) (noting citizens' contact with fair institutional procedures can support the legitimacy of disputed legal authorities during political transition).

[39] *See, e.g.*, Tom R. Tyler, *What is Procedural Justice? Criteria Used by Citizens to Assess the Fairness of Legal Procedures*, 22 LAW & SOC'Y REV. 103, 128–32 (1988).

[40] *See generally* Robert J. Bies & Debra L. Shapiro, *Voice and Justification: Their Influence on Procedural Fairness Judgements*, 31 ACAD. MGMT. J. 676 (1988).

[41] *See* JOHN W. THIBAUT & LAURENS WALKER, PROCEDURAL JUSTICE: A PSYCHOLOGICAL ANALYSIS 1–2 (1975).

[42] E. Allan Lind et al., *Voice, Control, and Procedural Justice: Instrumental and Noninstrumental Concerns in Fairness Judgments*, 59 J. PERSONALITY & SOC. PSYCH. 952, 952 (1990).

[43] *Id.*; Tom R. Tyler et al., *Influence of Voice on Satisfaction with Leaders: Exploring the Meaning of Process Control*, 48 J. PERSONALITY & SOC. PSYCH. 72, 77 (1985); *see also* Marco Kleine et al., *How Voice Shapes Reactions to Impartial Decision-Makers: An Experiment on Participation Procedures*, J. ECON. BEHAV. & ORG. 241, 241–42 (2017). *But see* Derek R. Avery & Miguel A. Quiñones, *Disentangling the Effects of Voice: The Incremental Roles of Opportunity, Behavior, and Instrumentality in Predicting Procedural Fairness*, 87 J. APPLIED PSYCH. 81, 81–82, 85 (2002) (distinguishing between voice opportunity and voice behavior and finding that "when voice instrumentality is low, voice behavior has a negative impact on procedural fairness").

[44] *See* Lind et al., *supra* note 42.

[45] Brockner et al., *Culture and Procedural Justice: The Influence of Power Distance on Reactions to Voice*, 37 J. EXPERIMENTAL SOC. PSYCH. 300, 312–13 (2001); Kees van den Bos et al., *The Consistency Rule and the Voice Effect: The Influence of Expectations on Procedural Fairness Judgements and Performance*, 26 EUR. J. SOC. PSYCH. 411, 423–26 (1996); David de Cremer & Jeroen Stouten, *When Does Giving Voice or Not Matter? Procedural Fairness Effects as a Function of Closeness of Reference Points*, 24 CURRENT PSYCH. 203, 210 (2005); *see* Joseph P. Daly & Paul D. Geyer, *The Role of Fairness in Implementing Large-Scale Change: Employee Evaluations of Process and Outcome in*

the value-expressive function of voice may sometimes be elevated above its other functions.[46] But for voice to convey respect and inclusion, individuals must feel listened to; they must experience their participation as meaningful and not merely a sham.[47]

People also tend to endorse a procedure as fair if decisions are openly justified.[48] By giving reasons, decisionmakers reassure the parties that they have "acted on the presented viewpoints in an impartial and unbiased manner."[49] Unsurprisingly, it is often losers rather than winners who demand justifications for outcomes. To satisfy them, the explanations have to be perceived as sincere and adequate.[50] More specific or thorough explanations will generally be seen as more adequate.[51]

Procedures believed to be fair may not in fact be so; but the distinction between descriptive and normative theories "does not force the conclusion that litigant satisfaction is unimportant or that it should not be considered in the evaluation and comparison of specific procedures." [52] Giving disputants satisfaction and closure is an essential aspect of any justice system, and Tim Wu identifies procedural fairness as an "obvious" advantage that human judges have over their artificial rivals.[53]

But is this advantage so obvious? When it comes to voice, advances in natural language technology have empowered computers to convert between speech and text, give increasingly intelligent replies to questions, summarize documents, and spot contradictions in statements.[54] These advances raise the

---

Seven Facility Relocations, 15 J. ORG. BEHAV. 623, 634 (1994); Patricia Grocke et al., *Young Children Are More Willing to Accept Group Decisions in Which They Have Had a Voice*, 166 J. EXPERIMENTAL CHILD PSYCH. 67, 75 (2018).

[46] Lind et al., *supra* note 42.

[47] Tom R. Tyler, *Conditions Leading to Value-Expressive Effects in Judgments of Procedural Justice: A Test of Four Models*, 52 J. PERSONALITY & SOC. PSYCH. 333, 339 (1987).

[48] *See, e.g.*, Robert J. Bies, *Beyond "Voice": The Influence of Decision-Maker Justification and Sincerity on Procedural Fairness Judgements*, 17 REPRESENTATIVE RSCH. SOC. PSYCH. 3, 10 (1987); Bies & Shapiro, *supra* note 40, at 683.

[49] Bies, *supra* note 48, at 4; *see also* Daly & Geyer, *supra* note 45, at 627.

[50] *See* Robert J. Bies et al., *Causal Accounts and Managing Organizational Conflict: Is It Enough to Say It's Not My Fault?*, 15 COMMC'N RSCH. 381 (1988) (studying excuses that employers gave for refusing their employees' requests); Daly & Geyer, *supra* note 45.

[51] *See* Debra L. Shapiro et al., *Explanations: What Factors Enhance Their Perceived Adequacy*, 58 ORG. BEHAV. & HUM. DEC. PROCESSES 346, 346 (1994); Debra L. Shapiro, *The Effects of Explanations on Negative Reactions to Deceit*, 36 ADMIN. SCI. Q. 614, 614 (1991); *see also* Tania Lombrozo, *Simplicity and Probability in Causal Explanation*, 55 COGNITIVE PSYCH. 232, 232 (2007) (noting that there is a distinction between normative justifications for a decision and causal explanations of a phenomenon; people favor simpler causal explanations over more complex ones).

[52] Lawrence B. Solum, *Procedural Justice*, 48 S. CAL. L. REV. 181, 266 (2004).

[53] Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2002–03 (2019).

[54] *See generally* Katja Grace et al., *Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts,* 62 J. A.I. RSCH. 729 (2018).

possibility that parties could one day have their grievances and pleas heard by machines, as though they were before human judges.

Now, the ability to perform the above-mentioned tasks does not imply that machines understand language like humans do. Even if machines could generate perfect sentences in Chinese, they do so by learning word frequencies and co-occurrences or by obeying a grammar logic they have been taught.[55] They don't actually know the meaning of the sentences they are parsing; they hear without comprehending and utter without intention. While this asserted difference between human and artificial minds seems founded on little more than intuition, recent challenges to established benchmarks in computational linguistics are telling. Dubbed "adversarial attacks," these evaluative tests undermine the notion that high-performing, state-of-the-art, algorithms have a semantic grasp of language. Machines adapt poorly to texts that are marginally different from those they have encountered before. Introducing ungrammatical distractors into passages, for example, reduces the accuracy for some algorithms from over 75% to a mere 7%.[56] So it is reasonable to think that a hearing before a machine may not be qualitatively the same as a hearing before a human.

At the same time, however, the issue of whether machines truly understand might be irrelevant to how humans respond to them. Computers are frequently depicted as static installations that are distant and inscrutable. But computers can also be portrayed as corporeal systems possessing the capacity for thought, emotion, and even humor — R2-D2 is an example. They are, on an influential theory, social actors.[57] Studies find that people tend to apply rules of social behavior to human-computer interactions despite recognizing the inapplicability of those same rules to machines.[58] We are gentler in rating a computer when the evaluation is requested by the computer itself rather than a human third party.[59] We are partial to "silicon sycophants" that flatter us.[60] We even project gender onto machines, heeding the advice of computers represented as male on "masculine" topics and computers represented as female on "feminine" topics.[61] The "computers as social actors"

---

[55] *See* John R. Searle, *Minds, Brains, and Programs*, 3 BEHAV. & BRAIN SCI. 417, 417–18 (1980).

[56] *See* Robin Jia & Percy Liang, *Adversarial Examples for Evaluating Reading Comprehension Systems*, 2017 PROC. CONF. ON EMPIRICAL METHODS NAT. LANGUAGE PROCESSING 2021, 2022.

[57] *See generally* Clifford Nass et al., *Computers Are Social Actors*, 1994 PROC. SIGCHI CONF. ON HUM. FACTORS COMPUTING SYS. 72 (arguing that individuals' interactions with computers are social).

[58] *See* Clifford Nass & Youngme Moon, *Machines and Mindlessness: Social Responses to Computers*, 56 J. SOC. ISSUES 81, 85 (2000); Youjeong Kim & S. Shyam Sundar, *Anthropomorphism of Computers: Is It Mindful or Mindless?*, 28 COMPUTS. HUM. BEHAV. 241, 241 (2012).

[59] *See* Nass et al., *supra* note 57, at 74.

[60] B.J. Fogg & Clifford Nass, *Silicon Sycophants: The Effects of Computers that Flatter*, 46 INT'L J. HUM.-COMPUT. STUD. 551, 552–52 (1977).

[61] *See* Eun-Ju Lee, *Effects of "Gender" of the Computer on Informational Social Influence: The Moderating Role of Task Type*, 58 INT'L J. HUM.-COMPUT. STUD. 347, 347–48 (2003).

paradigm posits that extant norms of procedural fairness will govern machine adjudication: People will rate an algorithm as fairer if they have an opportunity to speak to the robot deciding their cases.

As for justification, explanations matter in part because they help demonstrate the absence of judicial bias. But concerns about bias might be attenuated for machines. While "[t]he great tides and currents which engulf the rest of men do not turn aside in their course and pass the judges by,"[62] they may not sway computers. In fact, D'Amato speculated that:

> [L]aw might seem more impartial to the man on the street if computers were to take over large areas now assigned to judges. There is certainly some degree of belief on the part of the public that judges cannot escape their own biases and prejudices and cannot free themselves from their relatively privileged class position in society. But computers, *unless programmed to be biased*, will have no bias. They will give the same result on the same facts irrespective of the race, color, wealth, talents, or deference of the litigants.[63]

D'Amato's qualification is crucial: There is a nagging worry that algorithmic processes are perpetuating the same biases that infect humans.[64] Indeed, academics and popular writers have sounded the alarm about algorithms that discriminate.[65] Because AI is sometimes presented as a black box, there is little reassurance that machines are not taking protected characteristics into account, thereby reproducing invidious stereotypes.[66] Insofar as algorithms are trained on datasets of human decisions, this is not entirely surprising: "[B]ias in, bias out."[67] One natural solution is disclosure. To the extent people are suspicious of the factors and variables machines consider, transparency about inputs might assuage some fears.[68]

---

[62] BENJAMIN N. CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 168 (1921).

[63] D'Amato, *supra* note 18, at 1300 (emphasis added).

[64] *See, e.g.*, FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 35 (2015); Sandra Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218, 2221 (2019).

[65] *See, e.g.*, Claire Cain Miller, *When Algorithms Discriminate*, N.Y. TIMES (July 9, 2015), https://www.nytimes.com/2015/07/10/upshot/when-algorithms-discriminate.html [https://perma.cc/8CQD-9U2Y]; Rebecca Heilweil, *Why Algorithms Can Be Racist and Sexist*, VOX (Feb. 18, 2020), https://www.vox.com/recode/2020/2/18/21121286/algorithms-bias-discrimination-facial-recognition-transparency [https://perma.cc/LHT8-JG24].

[66] *See* Anya E.R. Prince & Daniel Schwartz, *Proxy Discrimination in the Age of Artificial Intelligence and Big Data*, 105 IOWA L. REV. 1257, 1275 (2020).

[67] Mayson, *supra* note 64, at 2224.

[68] *See* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathany & Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113, 152 (2018). As the authors note, however, giving algorithms access to protected characteristics may actually promote equity by enabling machines to learn the indicators that are actually predictive for a subgroup of the population. *See id.* at 154–60.

Secrecy, however, is not the only anxiety people have about algorithmic judging in general.[69] AI may also be opaque to users and even system designers themselves because the relationship between inputs and outputs is obscure and hard to fathom.[70] Clarity about the optimization function and the training data does not guarantee interpretability of the mechanism or its results.[71] Certainly, it is possible to furnish the parties a description of the computations being performed by their machine adjudicator. Intuiting the reasoning immanent in an algorithmic decision, however, often requires some sense of how the output conclusion might change given different input facts and circumstances.[72] Some machine learning techniques lend themselves readily to this kind of counterfactual thinking while others resist easy analysis. Tree-based methods, for example, are said to belong to the former category while deep neural network architectures fall into the latter. For this reason, proponents of interpretable artificial intelligence have recommended exploiting deep neural networks for their accuracy while rendering them explainable through an approximation by decision trees.[73]

Still, amidst the disquiet about AI, it must be kept in mind that humans are not always open and honest in their reasoning, either.[74] According to computer scientist Jon Kleinberg and coauthors, machines offer "far greater" visibility into "the ingredients and motivations of decisions, and hence far greater opportunity to ferret out discrimination."[75] "[T]here is instead every

---

[69] Andrew D. Selbst & Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085, 1087 (2018).

[70] *See, e.g.*, Davide Castelvecchi, *Can We Open the Black Box of AI?*, 538 NATURE 21, 21–22 (2016).

[71] *See*, *e.g.*, Brent Mittelstadt, Chris Russell & Sandra Wachter, *Explaining Explanations in AI*, PROC. CONF. ON FAIRNESS, ACCOUNTABILITY & TRANSPARENCY 279, 280 (2018) (explaining the distinction in the literature between "transparency" and "post-hoc interpretation"). For further discussion of interpretability in the machine learning community, see Zachary C. Lipton, *The Mythos of Model Interpretability*, 16 ACM QUEUE 31, 32 (2018); *see also* Doshi-Velez & Been Kim, *supra* note 16, at 9.

[72] *See, e.g.*, Sandra Wachter, Brent Mittelstadt & Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841, 844 (2018); Lara Kirfel & Alice Liefgreen, What If (and How…)? – Actionability Shapes People's Perceptions of Counterfactual Explanations in Automated Decision-Making 1 (2021) (unpublished manuscript) (on file with authors).

[73] *See* Leilani H. Gilpin et al., *Explaining Explanations: An Overview of Interpretability of Machine Learning*, 5 IEEE INT'L CONF. ON DATA SCI. & ADVANCED ANALYTICS 80, 82–83 (2018) (describing the use of proxy models to make deep neural architectures more explainable); *see also* Alwin Wan et al., *NBDT: Neural-Backed Decision Trees*, INT'L CONF. ON LEARNING REPRESENTATIONS 1 (2021) (proposing a hybrid between neutral networks and decision trees).

[74] *See* Kleinberg, Ludwig, Mullainathany & Sunstein, *supra* note 68, at 163; *see also* Joshua A. Kroll, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson & Harlan Yu, *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.").

[75] Kleinberg, Ludwig, Mullainathany & Sunstein, *supra* note 68, at 163.

reason to think," says Aziz Huq, that "[human] judicial discretion has had dismaying and socially destructive effects."[76]

II. TWO EXPERIMENTAL STUDIES

While there is rich theoretical literature on AI judging and legal processes, many of the key questions rest on open empirical claims about how people would evaluate AI judges. Perhaps the most immediate concern about AI judges is that ordinary citizens would see them as procedurally unfair, a harm in itself that also threatens public compliance with the law. This prompts the following questions:

1. Do ordinary citizens evaluate an AI-led judicial proceeding as less fair than a similar human-led one?
2. Do ordinary citizens evaluate an AI-led judicial proceeding as less fair than *any* human-led one?
3. Could an AI judge give an ordinary citizen a sense of being fairly heard?
4. When it comes to fairness, do people see the interpretability of decisions as more critical for AI judges than human judges?
5. Do people's assessments of the fairness of AI judges vary by legal contexts or issues? For example, are people more amenable to private law AI arbitrators compared to AI criminal law judges?

This Part presents two experimental studies of ordinary citizens, studies that offer some of the first empirical evidence bearing on each of these central and largely untested questions. All of the studies' materials, including pre-registrations, vignettes, and data have been made available online.[77]

A. Study 1

We investigate how people perceive the fairness of human as opposed to AI judges in three different adjudicatory contexts: consumer refund for a damaged product, pre-trial bail for criminal offenses, and custodial sentencing post-conviction. These contexts were presented in the form of vignettes that experimental subjects were asked to read and evaluate, which are presented below.

1. Experimental Scenarios

Consumer refund arbitration scenario. This scenario recounted a disagreement over the physical condition of goods sold and delivered by a merchant to a customer.

> *John Smith is 25 years old. Recently, John ordered a high-end camera for $2500 from an online retailer called "Camerazon."*

[76] Huq, *supra* note 9, at 666.

[77] *Having Your Day in Robot Court*, OPEN SCI. FRAMEWORK, https://osf.io/cw2m4 (last visited Oct. 26, 2022).

> *John paid for the camera with a credit card and selected a home delivery option. The next day, John received the camera in the mail. When he opened the package, he saw what he believed to be a small smudge on the camera lens. John tried to wipe the lens clean with a lens cloth, but the smudge did not disappear.*
> *The Camerazon policy states clearly that if the goods were delivered in a damaged state, Camerazon would refund the purchase. John emailed Camerazon's customer service and included a photo of the camera lens. A Camerazon representative denied the refund, stating that the goods do not appear to be damaged. John then sent several photos taken with the new camera, claiming that the mark was causing the photos to be discolored. Camerazon replied that they were sorry for John's dissatisfaction with the product, but that the photos taken did not appear to be discolored and thus they would not refund the purchase.*
> *Frustrated because he felt misled, John decided to pursue legal action against Camerazon. The purchase terms stated that all disputes must be resolved in arbitration.*
> *John filed an arbitration claim, seeking a refund for the camera, which John claimed was damaged. Both John and Camerazon agreed that, if the camera was damaged, he should be refunded. Moreover, they both agree that a permanent smudge that discolors photos would count as "damage" qualifying for refund. The dispute between the parties centered around:*
>
> > *(1) whether there was a smudge mark on the camera; and*
> > *(2) whether the photographs were discolored.*
>
> *The arbitration decision would be made on the basis of these two factors.*

Pre-trial bail scenario. This scenario concerned an arrest and prosecution for marijuana possession:

> *John Smith is 25 years old. Recently, the police discovered four pounds of marijuana in the trunk of John's car during a routine traffic stop. John was arrested. Because of the large amount of marijuana found, the prosecutor decided to charge John for possessing marijuana with the intent to distribute. John will be tried in court to determine whether he is guilty or innocent. If he is found guilty, he could face up to five years of imprisonment.*
> *However, even before trial, a decision has to be made whether to keep John in custody or to grant him bail. If the court decides to grant bail, John will have the opportunity to pay an amount of money to ensure his appearance at the trial. If he pays the bail amount, John will not be jailed before the trial.*

> *The bail amount will be refunded to John after the trial is over. If the court decides to keep John in custody, he will have to stay in jail until his trial starts. John will not be compensated for the time he spent in jail even if he is subsequently acquitted at trial.*
> *Anxious because he was his family's sole breadwinner, John asked the court for bail.*
> *There are two reasons that a court might decide to keep John in custody in this context:*
>
> > *(1) flight risk: the risk that John would flee before his trial; and*
> > *(2) further offenses risk: the risk that John might commit further criminal offenses before his trial.*
>
> *Indeed, the law requires bail determinations to be made primarily on the basis of these two risks but it does not dictate how these risks are to be assessed.*

Custodial sentencing scenario. The last scenario revolved around a case of manslaughter:

> *John Smith is 25 years old. Two years ago, John was laid off from his job. After being unemployed for a full year, John felt in desperate need of cash. He was not happy with his options, and ultimately he decided to rob a bank. John owned a gun that he used for recreational shooting at a local range. As he left for the robbery, he took the gun with him. He didn't intend to use it, but thought it might be useful.*
> *When John arrived at the bank, things did not go to plan. He demanded that the teller hand over all the cash in her register. The teller had been in very poor health recently, but John did not know this, as he had never before met the teller. John thought she was not acting quickly enough, so he took out the gun and waved it in front of her to speed things up. Seeing the gun, the teller was struck with fear and began to have a heart attack. She handed over a large stack of bills before collapsing from the heart attack. John fled with the money. Thirty minutes later, police arrived on the scene. But the bank teller's heart attack had already killed her.*
> *Eventually, the police tracked down John and arrested him. The state's prosecutor brought two charges against John, one for murder and one for manslaughter. The prosecutor made John a plea offer: If he pled guilty to the lesser offense of manslaughter, the prosecutor would drop the murder charge. John decided to take the deal. He pled guilty to manslaughter. Distressed because he did not intend the consequences of his actions, John asked the court for lighter punishment.*
> *Now, John is about to receive his sentence for manslaughter. In the state in which John was convicted, the sentencing guidelines for manslaughter indicate a mandatory minimum*

> *sentence of at least five years in prison and a maximum sentence of fifteen years.*
> *The sentencing factors include:*
> *(a) the nature of the crime,*
> *(b) the character and history of the defendant, such as whether John has a criminal history, and*
> *(c) whether John was under great personal stress or duress when committing the crime.*
> *The sentencing decision would be made on the basis of these factors but the law does not dictate how these factors are to be assessed.*

Subjects were randomly assigned to one of these three scenarios. In all three scenarios, the ultimate judicial decision went against John: he did not obtain a refund for the allegedly damaged camera, he was denied bail pending trial for possession of a controlled substance with intent to distribute, and he received the maximum possible sentence for manslaughter.

2. Experimental Treatments

The way these negative decisions were reached varied across three factors. First, we manipulated whether the decisionmaker was a human or an algorithm ("Agent"). Second, the decision could have been made with or without a hearing ("Hearing"). Third, the decision was interpretable or not interpretable ("Interpretability").

Human or AI judge. Because our primary aim was to assess differences in lay assessments of human and AI judges, the human and the algorithmic decisionmakers were introduced as being very competent at their adjudicative tasks. For example, in the pre-trial bail scenario, subjects randomized to the "human" condition were told that:

> *[i]n the state where John was arrested and charged, bail decisions are made by a judge. These judges are very experienced and can predict flight and further offenses risk to a very high degree of accuracy. Among other things, the judge already has information about John's background, his previous convictions, and potential extenuating circumstances if any.*

Similarly, those randomized to the "algorithm" condition read that

> *[i]n the state where John was arrested and charged, bail decisions are made by an algorithm. This algorithm employs advanced statistical and machine learning techniques and can predict flight and further offenses risk to a very high degree of accuracy. Among other things the algorithm already has information about John's background, his previous convictions, and potential extenuating circumstances if any.*

Hearing. Second, the decision could have been made based on the record only or after a hearing. For example, in the consumer refund case, subjects randomized to the "algorithm" and "hearing" conditions were informed that

> *[b]efore an algorithm makes a decision, sometimes there is an arbitration hearing, but sometimes there is not. In John's case, there is an arbitration hearing. John has an opportunity to present his case in person. The hearing allows John to explain why the camera was damaged and therefore should be refunded, by speaking to a computer that transcribes his speech for consideration by the algorithm. Through this hearing, the algorithm is able to evaluate John's credibility and emotions.*

The lack of a hearing was also made explicit. Thus, subjects randomized to the "human" and "no hearing" conditions were told that

> *[b]efore an arbitrator makes a decision, sometimes there is an arbitration hearing, but sometimes there is not. In John's case, there is not an arbitration hearing. John does not have an opportunity to present his case in person. The hearing would have allowed John to explain to the arbitrator why the camera was damaged and therefore should be refunded. Through this hearing, the arbitrator would have been able to evaluate John's credibility and emotions.*

Interpretability. Third, the decision could be interpretable, or not. Interpretability here does not refer to the articulation of reasons in support of a disposition. Rather, it entails transparency into—and knowledge of—how the outcome is derived. Thus, under the "interpretable" condition, the vignette concludes by stating that

> *[w]hile the [arbitrator|judge|algorithm]'s reasoning is rigorous, it is also easy to understand. All factors were considered using a flowchart that asks at each stage whether a particular criteria is satisfied. It would therefore be possible for John, or anyone else, to figure out how much each factor mattered to the [decisionmaker]'s ultimate decision. Moreover, it would be possible for someone else to replicate the [decisionmaker]'s reasoning to see how a change in any of his factors impacts the sentencing decision.*

In contrast, under the "uninterpretable" condition, the vignette ends by admitting that

> "[w]hile the [decisionmaker]'s reasoning is rigorous, it is also not easy to understand. All factors were considered, but given the complex nature of the decision-making process, it is not possible to describe in simple terms how the [decisionmaker] decision was produced."

3. Hypotheses

Given the tenor of procedural justice literature, we anticipate that decisions reached after a hearing will be judged as fairer than those rendered on the record. We also expect that decisions will be judged as fairer if they are interpretable rather than uninterpretable. But it remains unclear whether human adjudication will always have a perceived procedural fairness advantage over artificial intelligence. On the one hand, it seems almost axiomatic that some uniquely human qualities, such as empathy, are necessary for the parties to feel they have been listened to and given a fair shake. Moreover, people are not accustomed to having machines resolve their disputes, and computers — unlike humans — are vulnerable to hardware malfunction or programming bugs. Hence, there could be some uneasiness about having algorithms determine matters of great import.[78] On the other hand, people sometimes trust the advice of computers, believing them to be better at objective tasks than experts.[79] Rightly or wrongly, people also tend to conceive of machines as being rule-bound and, hence, less capricious than humans who may succumb to passion or preconception.[80]

The scenarios employed in our experiment differ in terms of the consequences and the adjudicative task. At stake in the refund decision is $2,500; in the bail decision, time spent in jail between committal and trial; and in the sentencing decision, a difference of ten years in prison between the lower and upper ends of the sentencing range. Moreover, the refund decision rests on "whether there was a smudge mark on the camera" and "whether the photographs were discolored," whereas the bail decision has to be made based on "flight risk" and "further offenses risk." The former set of variables aid in the classification of a physical object whereas the latter requires prediction of future behavior. No moral evaluation, however, is involved. In contrast, the sentencing decision has to account for "the nature of the crime," "the character and history of the defendant," and whether [the defendant] was under great personal stress or duress when committing the crime." The law thus calls for a normative balancing of several factors — considerations that bear on recidivism and rehabilitation but that also speak to blame and culpability. In sum, it is plausible that the use of AI for adjudication will be perceived as fairer when the issue is a refund for a damaged product, rather than sentencing for a manslaughter conviction; compared to the former scenario, the latter requires normative balancing and moral evaluation and also involves higher stakes.

---

[78] *See* Markus Langer, Cornelius J. König, & Maria Papathanasiou, *Highly automated job interviews, Acceptance under the influence of stakes*, 27 INT'L J. SELECTION & ASSESSMENT 217, 228 (2019).

[79] *See* Noah Castelo et al., *Task-Dependent Algorithm Aversion*, 56 J. MKTG. RSCH. 809, 8218 (2019); *see also* Jennifer M. Logg et al., *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment*, 151 ORG. BEHAV. & HUM. DEC. PROCESSES 90, 93 (2019).

[80] *Cf.* Natali Helberger, Theo Araujo & Claes H. de Vreese, *Who is the Fairest of Them All? Public Attitudes and Expectations Regarding Automated Decision-Making*, 39 COMPUT. L. & SEC. REV. 1, 9, 11 (2020) (noting that though there is less capricious decision-making by machines, which some also view that in an unfavorable light).

Finally, the contribution of voice and interpretability to procedural justice may depend on the cognitive and emotive capacities of the decisionmaker. The opportunity to speak and be heard might only be regarded as meaningful if the adjudicator is able not only to parse language but to truly understand and empathize with the parties. Humans, unlike algorithms, possess these abilities. Moreover, demands for transparency and insight into adjudicatory decision-making become more acute when there is a danger of outcomes being tainted by illicit motivations. Humans, unlike algorithms, might be motivated by their own interests and prejudices. Because algorithms have neither emotions nor desires, voice and interpretability might not enhance the perceived fairness of artificial justice. At the same time, people ascribe mental states to computers, projecting norms, beliefs, and stereotypes onto them. The human tendency to anthropomorphize machines implies that both voice and interpretability will continue to matter, even in the brave new world of AI adjudication.

The generalizability of basic findings in procedural justice research is tested by randomizing subjects to the consumer refund, pre-trial bail, or custodial sentencing scenario, the "human" or "algorithm" condition, the "hearing" or "no hearing" condition, and the "interpretable" or "not interpretable" condition. This first study thus features a between-subject, 3×2×2×2 factorial design, meaning that each participant reads a single vignette describing a randomly selected scenario featuring randomly varied facts.; For example, a participant might be randomly assigned to the consumer refund scenario, with an algorithmic decisionmaker that is not interpretable and a hearing.

Scenario	**Decisionmaker**	**Interpretability**	**Hearing**
Consumer Refund	Human	Interpretable	Hearing
Pre-trial Bail	Algorithm	Not Interpretable	No Hearing
Sentencing			

Table 1: Four Factors in the 3x2x2x2 Between-Subjects Design

Before reading any of the scenarios, subjects were first queried about their trust in legal authorities, measured on a 1 to 7 scale, 1 being "no trust" and 7 being "complete trust." They were then instructed to read their randomly assigned vignettes and surveyed for their reactions. Specifically, subjects were invited to rate, on a 1 to 7 scale—1 being "strongly disagree" and 7 being "strongly agree"—whether they agreed or disagreed that the decisionmaker's procedure for arriving at the decision was fair, whether the decisionmaker considered all relevant facts in making the decision, and whether the decisionmaker understood John's perspective in making the decision. These statements were displayed on separate pages. Subjects were also requested to estimate, from 0 to 100, 0 being "incorrect every time," and 100 being "correct every time," how accurate they believed the decisionmaker to be in making

decisions. The final item in the section asked subjects whether they thought John felt he had a good opportunity to voice his own arguments about the decision. Their responses were captured on a 1 to 7 scale, 1 being "definitely no," and 7 being "definitely yes."

To summarize, six variables were collected in this section of the protocol. In order, they are "Trust in Legal Authorities," "Procedural Fairness," "Thoroughness," "Understanding," "Accuracy," and "Voice." Manipulation check questions were posed at the end.

4. Data and Analysis

The experiment was fielded on 1,710 subjects in September 2020. These subjects were recruited through Lucid Theorem and sampled to be nationally representative of the United States population.[81] As a preliminary matter, the experimental manipulations were successful. 78.1% and 76.1% of subjects randomly assigned to the "hearing" and "no hearing" conditions respectively correctly recalled whether John had the chance to speak and have his credibility and emotions evaluated by the decisionmaker. In addition, 87.9% and 86.1% of subjects randomly assigned to the "interpretable" and "not interpretable" conditions correctly recalled how the decision was reached.

Pooling across all three scenarios and other factors, we find that substituting an algorithm for a human significantly diminished subjective judgments of procedural fairness (see Figure 1). Subjects assigned to the "algorithm" condition gave ratings that were on average 0.466 lower (p<0.001, two-sided t-test; p=0.002, two-sided Wilcoxon rank sum test) than those in the "human" baseline.[82]

On the other hand, the opportunity for a hearing and the interpretability of the decision had positive and significant effects on subjects' perceptions of the fairness of the adjudicative process (see Figure 2). Compared to the "no hearing" baseline, subjects in the "hearing" condition gave fairness ratings that were on average 0.297 higher (p=0.002, two-sided t-test; p=0.003, two-sided Wilcoxon rank sum test).[83] Compared to the "not interpretable" baseline, subjects in the "interpretable" condition gave fairness ratings that were on average 0.305 higher (p=0.002, two-sided t-test; p<0.001, two-sided Wilcoxon rank sum test).[84]

Overall, the direction and size of these effects do not appear to vary by scenario. In general, we examine the moderation of treatment effects by estimating an ordinary least squares regression model of the form

$$y = \beta_0 + \beta_1 I + \beta_2 T + \beta_3 I \times T$$

[81] Lucid Theorem is a service that helps recruit respondents for online studies. Lucid provides nationally representative samples of the United States population by quota sampling along the dimensions of age, gender, race, and politics.

[82] The Neyman estimator gives a conservative standard error of 0.097.

[83] The Neyman estimator gives a conservative standard error of 0.097. This estimate is equivalent to the HC2 robust standard errors estimated from a regression of the outcome variable on a treatment indicator.

[84] The Neyman estimator gives a conservative standard error of 0.097.

where y is the outcome of interest, I is an indicator variable for the moderator, and T is an indicator variable for the treatment. Then the effect in the absence of the moderator is β_2 while the effect in the presence of the moderator is $\beta_2 + \beta_3$. β_3—the coefficient on the interaction term $I \times T$—captures moderation in the treatment effect.[85] A linear regression of procedural fairness ratings on scenario indicators, a hearing indicator, and scenario-hearing interactions returns statistically insignificant coefficients for the bail-hearing (-0.250, p=0.298) and sentencing-hearing (-0.171, p=0.458) interaction terms. [86] Similarly, a linear regression of procedural fairness ratings on scenario indicators, an interpretability indicator, and scenario-interpretability interactions returns statistically insignificant coefficients for the bail-interpretability (0.907, p=0.686) and sentencing-interpretability (0.180, p=0.437) interaction terms.[87] A linear regression of procedural fairness ratings on scenario indicators, a decisionmaker indicator, and scenario-decisionmaker interactions also returns statistically insignificant coefficients for the bail-algorithm (-0.053, p=0.827) and sentencing-algorithm (-0.203, p=0.380) interaction terms.[88]

---

[85] *See* ANDREW GELMAN, JENNIFER HILL, & AKI VEHTARI, REGRESSION AND OTHER STORIES 134—38 (2021); *see* Jiannan Lu, *On randomization-based and regression-based inferences for 2^K factorial designs*, 112 STAT. & PROBABILITY LETTERS 72, 75—76 (2016).

[86] The estimated model is $y = \beta_0 + \beta_1 I_{Bail} + \beta_2 I_{Sentencing} + \beta_3 T_{hearing} + \beta_4 T_{hearing} \times I_{Bail} + \beta_5 T_{hearing} \times I_{Sentencing}$ where y is procedural fairness ratings, I_{Bail} is an indicator variable for the bail scenario, $I_{Sentencing}$ is an indicator variable for the sentencing scenario, and $T_{hearing}$ is an indicator variable for the presence of a hearing. Note that the reference levels are "refund" and "no hearing."

[87] The estimated model is $y = \beta_0 + \beta_1 I_{Bail} + \beta_2 I_{Sentencing} + \beta_3 T_{interpretable} + \beta_4 T_{interpretable} \times I_{Bail} + \beta_5 T_{interpretable} \times I_{Sentencing}$ where y is procedural fairness ratings, I_{Bail} is an indicator variable for the bail scenario, $I_{Sentencing}$ is an indicator variable for the sentencing scenario, and $T_{interpretable}$ is an indicator variable for the interpretability of the decision. Note that the reference levels are "refund" and "not interpretable."

[88] The estimated model is $y = \beta_0 + \beta_1 I_{Bail} + \beta_2 I_{Sentencing} + \beta_3 T_{algorithm} + \beta_4 T_{algorithm} \times I_{Bail} + \beta_5 T_{algorithm} \times I_{Sentencing}$ where y is procedural fairness ratings, I_{Bail} is an indicator variable for the bail scenario, $I_{Sentencing}$ is an indicator variable for the sentencing scenario, and $T_{algorithm}$ is an indicator variable for algorithmic decisionmaker. Note that reference levels are "refund" and "human."

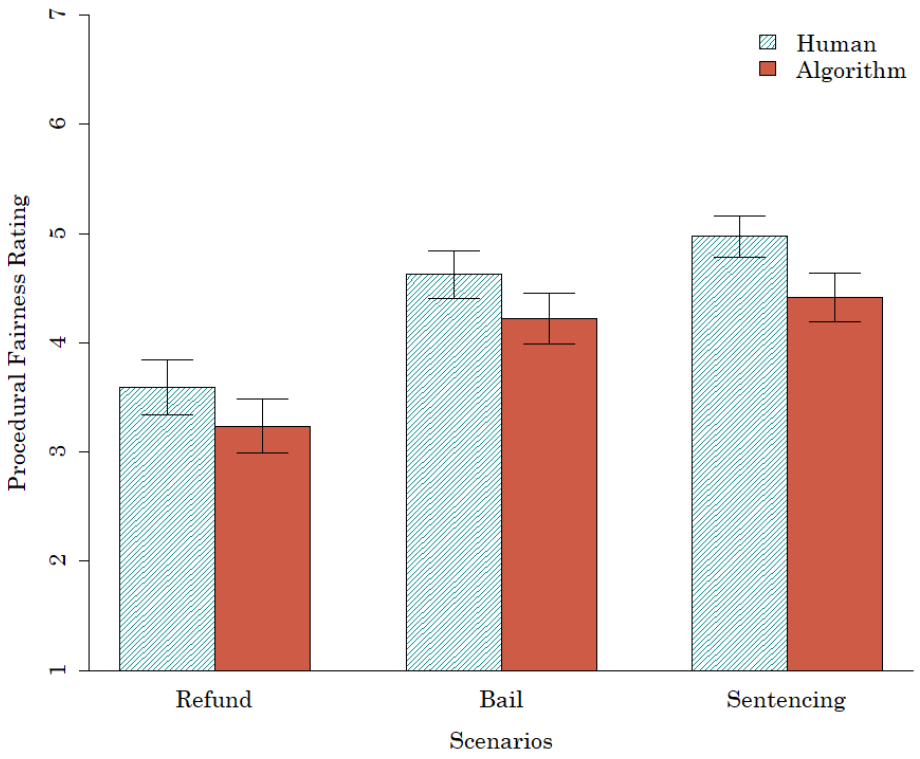

Figure 1, The Human-AI Fairness Gap: Average Procedural Fairness Rating in Study 1 by Scenario and Decisionmaker

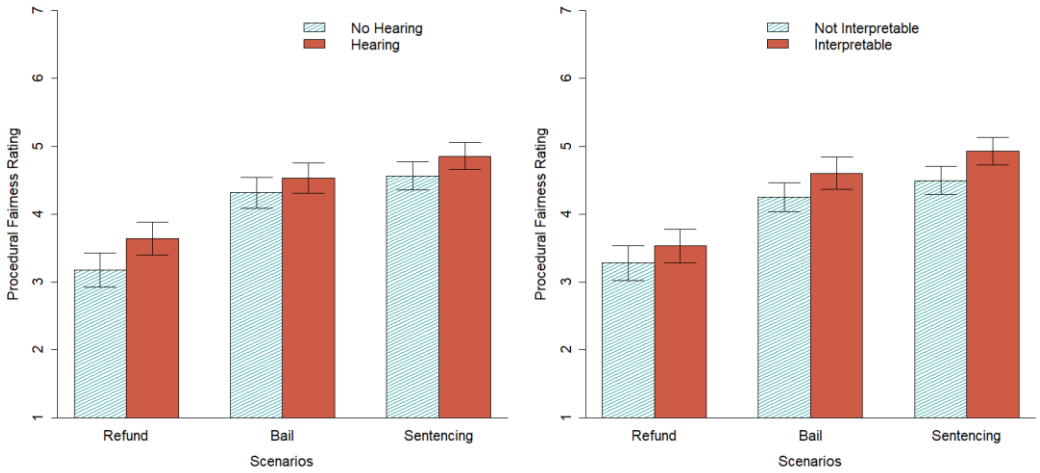

Figure 2, Hearing and Interpretability Increase Fairness: Average Procedural Fairness Rating in Study 1 by Scenario and Hearing/ Interpretability

	Hearing by Scenario	Interpretable by Scenario	Decisionmaker by Scenario

Constant	3.1753***	3.2780***	3.5920***
	(0.1279)	(0.1286)	(0.1271)
Hearing	0.4632**		
	(0.1775)		
Interpretable		0.2516	
		(0.1783)	
Algorithmic Decisionmaker			-0.3545*
			(0.1779)
Bail	1.1414***	0.9746***	1.0344***
	(0.1719)	(0.1685)	(0.1683)
Sentencing	1.3880***	1.2170***	1.3815***
	(0.1659)	(0.1661)	(0.1591)
Hearing: Bail	-0.2503		
	(0.2401)		
Hearing: Sentencing	-0.1714		
	(0.2307)		
Interpretable: Bail		0.0973	
		(0.2408)	
Interpretable: Sentencing		0.1795	
		(0.2310)	
Algorithmic Decisionmaker: Bail			-0.0525
			(0.2400)
Algorithmic Decisionmaker: Sentencing			-0.2027
			(0.2307)
Observations	1645	1645	1645
R^2	0.0842	0.0851	0.0901
Adjusted R^2	0.0814	0.0823	0.0874

Note: $^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$

Table 2: Estimated coefficients for Study 1 from an ordinary least squares regression of procedural fairness rating on indicator variables for scenario, an indicator variable for hearing/interpretability/type of decisionmaker, and the interaction between the variables. Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses.

We also investigate whether the influence of a hearing and interpretability on procedural justice judgments varies by the type of decisionmaker. To do so, we linearly regress procedural fairness ratings on a decisionmaker indicator, a hearing indicator, and the interaction between both indicators (Table 3).[89] The estimate for coefficient of the interaction term is negative, though it falls short of conventional levels of statistical significance (-0.337, p=0.080). We also linearly regress procedural fairness ratings on a decisionmaker indicator, an interpretability indicator, and the interaction between both indicators (Table 3).[90] The estimate for coefficient of the interaction term is positive but statistically insignificant (0.243, p=0.207).

To summarize, consistent with the prior literature on human decisionmakers, we find that a hearing and interpretability do affect how people judge the procedural fairness of legal decisions. We also find that the type of decisionmaker matters. A decision made by an algorithm is viewed as less procedurally fair than a decision made by a human. The data appears to suggest that a hearing is more important than interpretability for perceived fairness when the decisionmaker is a human as opposed to when the decision is made by an algorithm. But these differences are statistically insignificant.

	Hearing by Decisionmaker	Interpretability by Decisionmaker
Constant	4.2104	4.3469
	(0.0976)	(0.0917)
Algorithmic Decisionmaker	-0.3033*	-0.5901***
	(0.1391)	(0.1340)
Hearing	0.4712***	
	(0.1320)	
Interpretability		0.1887
		(0.1333)
Algorithmic Decisionmaker: Hearing	-0.3369	
	(0.1926)	
Algorithmic Decisionmaker: Interpretability		0.2434
		(0.1929)
Observations	1645	1645
R^2	0.0216	0.0213
Adjusted R^2	0.0198	0.0192
Note:	*p<0.05 **p<0.01 ***p<0.001	

[89] The estimated model is $y = \beta_0 + \beta_1 T_{algorithm} + \beta_2 T_{hearing} + \beta_3 T_{algorithm} \times T_{hearing}$ where y is procedural fairness ratings, $T_{algorithm}$ is an indicator variable for algorithmic decisionmaker, and $T_{hearing}$ is an indicator variable for hearing.

[90] The estimated model is $y = \beta_0 + \beta_1 T_{algorithm} + \beta_2 T_{interpretable} + \beta_3 T_{algorithm} \times T_{interpretable}$ where y is procedural fairness ratings, $T_{algorithm}$ is an indicator variable for algorithmic decisionmaker, and $T_{interpretable}$ is an indicator variable for the interpretability of the decision.

Table 3: Estimated coefficients for Study 1 from an ordinary least squares regression of procedural fairness rating on an indicator variable for the type of decisionmaker, an indicator variable for hearing/interpretability, and the interaction of both variables. Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses.

B. Study 2

1. Scenario, Experimental Treatments, and Hypotheses

To probe any possible interaction between the hearing and interpretability factors and the type of decisionmaker, we replicate the first study, this time limiting the scenarios to pre-trial bail and employing a larger sample size. The pre-trial bail scenario was chosen because it involved moderate stakes and a not too normatively laden evaluative task. Simulations based on data from the first study indicate that an experiment conducted on 5000 subjects has 80% power to detect the interaction between hearing and type of decisionmaker and 75% power to detect the interaction between interpretability and type of decisionmaker. The experimental treatments remain the same. The second study thus features a 2×2×2 factorial design: the decisionmaker may be a "human" or "algorithm," there may be a "hearing" or "no hearing", and the decision may be "interpretable" or "not interpretable".

Decisionmaker	**Interpretability**	**Hearing**
Human	Interpretable	Hearing
Algorithm	Not Interpretable	No Hearing

Table 4: Three Factors in the 2x2x2 Between-Subjects Design

As before, the instrument collected data on six variables: The principal outcome of interest, "Procedural Fairness," as well as "Trust in Legal Authorities," "Thoroughness," "Understanding," "Accuracy," and "Voice". Manipulation checks were also performed at the end.

2. Data and Analysis

In March 2021, 5086 subjects were recruited through Lucid Theorem for the experiment. Once again, the experimental manipulations were successful. 81.0% and 77.7% of subjects randomly assigned to the "hearing" and "no hearing" conditions, respectively, correctly recalled whether John had the chance to speak and have his credibility and emotions evaluated by the decisionmaker. Moreover, 86.6% and 87.5% of subjects randomly assigned to the "interpretable" and "not interpretable" conditions correctly recalled how the decision was arrived at.

Confirming the results of the first study, the opportunity for a hearing and the interpretability of the decision had positive and significant effects on

subjects' perceptions of the fairness of the adjudicative process. The estimates here are very similar to those from the first study. Compared to the "no hearing" baseline, subjects in the "hearing" condition gave fairness ratings that were on average 0.287 higher (p<0.001, two-sided t-test; p<0.001, two-sided Wilcoxon rank sum test).[91] Compared to the "not interpretable" baseline, subjects in the "interpretable" condition gave fairness ratings that were on average 0.295 higher (p<0.001, two-sided t-test; p<0.001, two-sided Wilcoxon rank sum test).[92] Substituting an algorithm for a human, on the other hand, significantly lowered subjective judgments of procedural fairness. Subjects assigned to the "algorithm" condition gave ratings that were on average 0.578 lower (p<0.001, two-sided t-test; p<0.001, two-sided Wilcoxon rank sum test) than those in the "human" baseline.[93]

The second study did not find evidence for an interaction between hearing or interpretability and the type of decisionmaker A hearing increased ratings of fairness for both human and algorithmic decisionmakers and no difference in effect was detected across the two conditions (Table 5). Likewise, interpretability boosted subjective judgments of procedural justice for both human and algorithmic decisionmakers but again, no difference in effect was detected (Table 5).

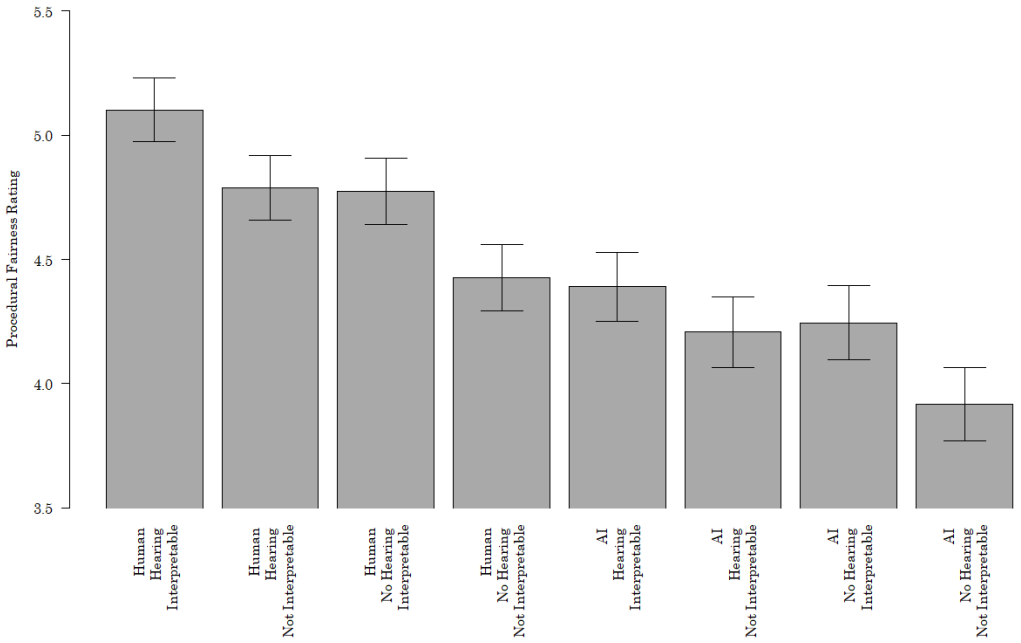

Figure 3: Average Procedural Fairness Rating in Study 2 by Decisionmaker, Hearing, and Interpretability. Note: To facilitate comparison of ratings across

[91] The Neyman estimator gives a conservative standard error of 0.051.
[92] The Neyman estimator gives a conservative standard error of 0.051.
[93] The Neyman estimator gives a conservative standard error of 0.050.

cells, the Figure's y-axis begins from 3.5 and ends at 5.5. The question presented a 1-7 scale.

	Hearing by Decisionmaker	Interpretability by Decisionmaker
Constant	4.592***	4.602***
	(0.048)	(0.048)
Algorithmic Decisionmaker	-0.516***	-0.541***
	(0.072)	(0.071)
Hearing	0.355***	
	(0.067)	
Interpretability		0.342***
		(0.067)
Algorithmic Decisionmaker: Hearing	-0.129	
	(0.100)	
Algorithmic Decisionmaker: Interpretability		-0.081
		(0.100)
Observations	5010	5010
R^2	0.033	0.033
Adjusted R^2	0.032	0.033
Note:	*p<0.05 **p<0.01 ***p<0.001	

Table 5: Estimated coefficients for Study 2 from an ordinary least squares regression of procedural fairness rating on an indicator variable for the type of decisionmaker, an indicator variable for hearing/interpretability, and the interaction of both variables. Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses.

3. Accounting for the Perceived Fairness Gap between Human and Algorithmic Decisionmakers

What accounts for the perceived procedural justice advantage of humans over algorithms demonstrated in the two studies presented here? There are several plausible explanations. Human judging may give the defendant an enhanced feeling of voice than algorithmic judging, even when there is no hearing. Relatedly, a human may be able to understand the defendant's situation in ways that an algorithm cannot. Alternatively, people may perceive a human as being more thorough or accurate than an algorithm.

To explore these possibilities, we collected information on potential mediator variables, namely, "Voice," "Understanding," "Thoroughness," and "Accuracy." Recall that the "Voice," "Understanding" and "Thoroughness" variables take on values between 1 and 7. A value of 1 indicates respectively

that the subject strongly disagreed that John felt that he had a good opportunity to voice his own arguments about the decision, that the decisionmaker understood John's perspective in making the decision, and that the decisionmaker considered all relevant facts in making the decision. A score of 7 indicates that the subject strongly agreed with these statements. "Accuracy," on the other hand, takes on values between 0 and 100, and represents the subject's estimate of the percentage of correct decisions rendered by the decisionmaker.

A variable is said to mediate an effect if the experimental treatment changes outcomes by changing the value of the "mediator." Take, for example, "Accuracy." If this variable fully mediates the effect of decisionmaker type on judgments of procedural fairness, then subjects who share the same estimate of the accuracy of the decisionmaker will rate the bail proceedings as equally fair whether it is administered by a human or an algorithm. To the extent that an algorithmic process is rated as less procedurally fair than one conducted by a human, it is because algorithms are perceived as less accurate than their human counterparts. In this case, we say there is no direct effect; the observed difference is entirely accounted for by the causal mediation effect in this example.[94]

Randomization of treatment alone is not sufficient for the identification and estimation of average direct and average causal mediation effects. Also required — but not usually feasible — is for the values of the mediator to be randomly set to the values they would assume under treatment or control. Concretely, if the candidate mediator were, say, "Voice," we would not only have to randomly assign subjects to the scenario where a human is the decisionmaker or the scenario where an algorithm is a decisionmaker. We would also have to manipulate subjects' beliefs about whether John felt he had a good opportunity to voice his arguments. And we would have to do so very

---

[94] More rigorously, let $Y_i(t, m)$ be the potential outcome if the treatment status were equal to t and the mediator variable took on the value m and let $M_i(t)$ denote the potential value of the mediator variable if the treatment status were equal to t. Then, the observed outcome for individual i is $Y_i(T_i, M_i(T_i))$. If individual i were assigned to receive treatment, then her outcome would be $Y_i(1, M_i(1))$; otherwise, it would be $Y_i(0, M_i(0))$. The treatment effect can thus be decomposed in the following way:
$Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = Y_i(1, M_i(t)) - Y_i(0, M_i(t)) + Y_i(1 - t, M_i(1)) - Y_i(1 - t, M_i(0))$
where $\xi_i(t) = Y_i(1, M_i(t)) - Y_i(0, M_i(t))$ is defined as the direct effect and $\delta_i(t) = Y_i(t, M_i(1)) - Y_i(t, M_i(0))$ as the causal mediation effect. The average direct effect $\bar{\xi}(t)$and average causal mediation effect $\bar{\delta}(t)$ are defined as the population averages of $\xi_i(t)$ and $\delta_i(t)$ respectively. Kosuke Imai et al., *Identification, Inference, and Sensitivity Analysis for Causal Mediation Effects*, 25 STAT. SCI. 51, 54 (2010). These notations and definitions may be extended to the case where there are multiple candidate mediators. Let $W_i(t)$ denote the potential value of the alternate mediators if the treatment status were equal to t, and let $M_i(t, w)$ and $Y_i(t, m, w)$ be the potential value of the mediator of interest and the potential outcome respectively. Then the causal mediation effect can be defined as $\delta_i(t) = Y_i\left(t, M_i(1, W_i(1)), W_i(t)\right) - Y_i\left(t, M_i(0, W_i(0)), W_i(t)\right)$. Kosuke Imai & Teppei Yamamoto, *Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments*, 21 POL. ANALYSIS 141, 147–49 (2013).

precisely — subjects' beliefs would have to be either the beliefs they would have had were the decisionmaker a human or the beliefs they would have had were the decisionmaker an algorithm. In the absence of such technology, causal mediation effects can only be isolated by making certain assumptions. In particular, we assume that procedural fairness ratings are statistically independent of the candidate mediators, conditional on the type of decisionmaker and pre-existing attributes of the subjects.[95] This assumption is strong and cannot be empirically verified. It may be tested by asking whether subjects' characteristics or dispositions might affect both the candidate mediators and the outcome variable. "Trust in Legal Authorities" is one such attribute. Subjects who place very little trust in legal authorities are likely to view the adjudicative process as unfair; they are also unlikely to believe that the judge — or algorithm — failed to consider all the facts or failed to understand the perspective of the defendant. We therefore adjust for "Trust in Legal Authorities" in the mediation analysis. Finally, we do not take causal independence between the putative mediators for granted. Subjects who consider that the decisionmaker understood John's perspective might, for that reason, also hold the opinion that the decisionmaker considered all the facts in arriving at the outcome. We make the necessary further assumption for causal mediation effects to be point-identified.[96]

Average causal mediation effects were computed for "Voice," "Understanding," "Thoroughness," and "Accuracy" using the *mediation* package for R.[97] A varying coefficient linear structural equations model was estimated for each candidate mediator, with the others posited as alternate mediators. This analysis indicates that only 2.0% of the reduction in fairness ratings that comes from having an algorithm rather than a human decide on bail is mediated by "Voice." The contributions of "Understanding," "Thoroughness," and "Accuracy," are 12.0%, 27.3%, and 29.3%, respectively.[98]

Candidate Mediator	Voice	Understanding	Thoroughness	Accuracy
Average Causal Mediation Effect	-0.012 (-0.019, 0.00)	-0.0714 (-0.092, -0.05)	-0.162 (-0.198, -0.13)	-0.174 (-0.209, -0.14)
Average Direct Effect	-0.582 (-0.673, -0.49)	-0.5230 (-0.609, -0.44)	-0.433 (-0.511, -0.35)	-0.419 (-0.499, -0.34)

[95] By random assignment, assignment to the experimental conditions is statistically independent of potential outcomes and candidate mediators. *See* Imai & Yamamoto, *supra* note 94, at 146—47.

[96] Specifically, we make the homogenous interaction assumption, i.e., $Y_i\big(1, m, W_i(1)\big) - Y_i\big(0, m, W_i(0)\big) = B_i + Cm$ for any m. *Id.* at 159.

[97] *Id.* at 158; *see also* Dustin Tingley et al., *Mediation: R Package for Causal Mediation Analysis*, 59 J. STAT. SOFTWARE 1, 26 (2014).

[98] These percentages are calculated by dividing average causal mediator effects by the total effect.

Total Effect	-0.594 (-0.685, -0.50)

Table 6: Estimates from varying coefficient linear structural equations model with procedural fairness ratings as the outcome variable, the type of decisionmaker as the treatment variable, the candidate mediator as the primary mediator, and the other mediators as alternate mediators. 95% confidence intervals are computed by bootstrap and reported in parentheses.

III. IMPLICATIONS

This Part discusses the implications of the experimental results. The first is a challenge for advocates of robot judges. Our study reveals that people generally see robot judges as less procedurally fair than human judges across different scenarios. Although others have raised concerns about algorithmic judging on the grounds of procedural justice, our study provides empirical data that is foundational to such a critique. In other words, although some scholars may not be surprised by the human-AI fairness gap, we offer some crucial empirical evidence to back up this claim.

These findings raise an objection to robot judges grounded in concerns about perceived fairness that go beyond any doctrinal objections. They also support a challenge grounded in (non-)compliance. Findings in legal psychology suggest that the legitimacy of the judicial system suffers if people see proceedings as unfair.

At the same time, the empirical results reveal a possible — and surprising — approach for making robot judging more acceptable to disputants. The study finds that lay perceptions of procedural fairness are also affected by the presence of a hearing and by the interpretability of the judge. Importantly, these factors increase the perceived fairness of *both* human and AI judges. In fact, we find that adding a hearing does not increase the perceived fairness of human-presided proceedings more than it does for AI- presided proceedings. Put simply, we don't find support for the intuition that people would find a hearing in front of an AI judge meaningless. We also find that people care about the interpretability of both human and AI decisions, calling into question the notion that ordinary citizens see human adjudication as, by it nature, more familiar or graspable than machine adjudication.

Strikingly, we also find that the decisionmaker, the hearing, and the interpretability has no stronger effect on procedural justice perceptions for high-stakes hearings in which the decision turns on ascribing a mental state to the defendant (sentencing) compared to low-stakes hearings that turn on factual determinations (consumer arbitration). This is surprising since one might have surmised the demands of procedural justice to be more stringent in the former scenario than in the latter. Moreover, mediation analysis suggests that the human-AI fairness gap is driven more by hard factors, like differing perceptions of accuracy, than "soft" and more distinctively human factors, like having one's voice heard. Together these results suggest that there may not be anything distinctive about human judges that prevents robot judges from closing the fairness gap.

Our findings suggest that the perceived fairness gap might indeed be closed through algorithmic offsetting, that is, by incorporating traditional elements of procedural justice into a machine-run adjudicative process. Enhancing the interpretability of an AI judge's decision, for instance, or allowing for a hearing before an AI judge could "offset" any perceived procedural justice penalty algorithms suffer vis-à-vis humans. Not all human judicial decisions are highly interpretable; nor do all human-led judicial proceedings involve a hearing. AI judges may be cheaper than human ones, and it may also be less costly or more feasible to increase the interpretability of or provide hearings before AI judges. The empirics presented in this Article indicate that, all else equal, proceedings before a human judge may be seen as no fairer than those conducted by AI judges that issue interpretable decisions after a hearing. Moreover, our data suggests that the more accurate algorithmic decision-making is thought to be, the fairer AI judging will be seen to be.

In the third Part below we address objections to our arguments, as well as limitations of the study. Other factors may affect people's evaluation of the fairness of judges (e.g., bias, accuracy), but our study suggests that these factors are not necessarily unique advantages of human judges. Moreover, any quality of human judges will vary from judge to judge; for example, human judges may reflect implicit racial bias, and some will exhibit more bias than others.

A. The Human-AI Fairness Gap: A Challenge for Robot Judges

Recall Chief Justice Roberts' answer about the timeliness of AI and judicial decision-making: "It's a day that's here."[99] Automated processes are already deployed in U.S. administrative practice,[100] and internationally, robot judges may soon become a reality.[101]

Whether robot judges can gain public acceptance is, however, a matter of contention. Of course, we cannot expect robot judges to be perfectly fair. Even human judges may fall short of such an ideal standard. In Eugene Volokh's words,"[o]ur question should not be whether AI judges are perfectly fair, only whether they are at least as fair as human judges."[102]

Our study assesses this exact question. Matched experimental scenarios manipulate the decision agent (human or algorithm) to assess whether Americans see robot-led proceedings as more unfair than human-led proceedings. We find a perceived fairness gap; human judges are seen as fairer than AI judges.[103] Moreover, this gap arises consistently across three distinct contexts: bail, sentencing, and commercial arbitration.

---

[99] Liptak, *supra* note 1.

[100] *See* Citron, *supra* note 4, at 1263–67.

[101] *See, e.g.*, Niiler, *supra* note 6 (discussing how Estonia plans on employing AI programs to decide certain small-claims cases).

[102] Volokh, *supra* note 17.

[103] These findings are consistent with prior experimental research on close, but ultimately distinct, questions. For example, Professor Simmons studies how people perceive judges that rely on algorithms as judicial aids. His study finds that people are

This discovery raises critical challenges for advocates of robot judges and governments preparing to implement them. For one, there may be good reason to care about people's understanding and evaluation of judicial fairness as an end in itself. For example, consider that our participants evaluated a process that lacked a hearing as less fair than a process that afforded one. This judgment might, in itself, be taken to provide a reason for our judicial system to offer opportunities for hearings. Of course, this reason is not decisive and might be outweighed by others. Perhaps it is prohibitively expensive for every type of adjudicative proceeding to include hearings. The point, however, that the human-AI fairness gap could be a reason — and not necessarily a decisive one — for adjudicative proceedings to employ human rather than robot judges.

Beyond this ethical argument, the results substantiate a legal compliance worry about robot judging. Legal psychologists hold that there is a relationship between perceived fairness and legal compliance.[104] If people regard robot judges as less fair, they may be less inclined to follow the laws that those robot judges would be charged to uphold. Introducing AI robot judges to reduce judicial administrative costs might come at a price of increased non-compliance.

The human-AI fairness gap—a major finding of our study—thus poses both ethical and legal compliance difficulties for having algorithms decide cases. At the same time, the other results from our study imply that humans may not have distinctive or absolute fairness advantages over machines.[105] If so, it might perhaps be possible to offset the fairness gap by affording more procedure in algorithm-presided proceedings. The next section develops this possibility.

B. Offsetting the Human-AI Fairness Gap

Beyond the main effect of the agent (human versus algorithm) on fairness, the experiment uncovered several other effects. Both interpretability and hearing improved judgments of procedural fairness. The tenor of these results is consistent with earlier research on procedural justice, conducted on human decisionmakers. For a human judge, more interpretable decisions were seen as fairer, and adding a hearing increased the perceived fairness of the proceeding.

One striking and novel finding in our study is that these same effects were observed for robot judges. That is, a hearing before a robot judge increased the perceived fairness of algorithmic adjudication; and more interpretable machine decisions were seen as fairer. Moreover, the effect on perceived fairness of adding a hearing was not appreciably larger for humans than for robot judges.

It is similarly striking that the human-AI fairness gap did not vary across scenarios. In the consumer arbitration scenario, the stake was $2,500, in the

---

skeptical of judges that use predictive algorithms. *See* Ric Simmons, *Big Data, Machine Judges, and the Legitimacy of the Criminal Justice System,* 52 U.C. DAVIS L. REV. 1067, 1108–09 (2018).

[104] *See* TOM R. TYLER, WHY PEOPLE OBEY THE LAW 3–4 (2006).

[105] *See supra* Figure 3 (demonstrating no statistically significant difference in average procedural fairness ratings between a human and an AI decisionmaker when the former makes uninterpretable decisions without a hearing and the latter renders interpretable decisions after a hearing).

sentencing scenario, the stake was ten years in. The dispositive issue in the consumer arbitration scenario was the determination of an objective fact, whereas the judge in the sentencing scenario had to ascertain the mental state of the criminal defendant. Conceivably, human advantages, if any, should have strengthen as the stakes increased and the issue in question went beyond factfinding. But we find no evidence of this.

These results complement the main finding concerning the human-AI fairness gap. Although humans are seen as fairer judges than robots, participants also evaluated those decision makers in a surprisingly similar way. That is, the perceived procedural fairness benefit of a hearing or an interpretable decision is not reserved solely for human judges. And we do not find that there are irreducible perceived fairness advantages of human decision makers even in a scenario as sensitive and consequential as sentencing. To illustrate, in our study, a human-adjudicated process with no hearing opportunity and resulting in an uninterpretable sentence was seen as no fairer than an algorithm-adjudicated process with the opportunity for a hearing and ending in an interpretable sentence (Figure 3).

These findings concerning interpretability and hearing raise the possibility of closing the perceived fairness gap through algorithmic offsettingAlgorithmic offsetting is possible insofar as human judges are not perceived as having a distinctive procedural justice advantage and to the extent that the features conducive to procedural fairness can be built into algorithmic adjudication. In our study, those features include the addition of a hearing, greater interpretability and, perhaps, accuracy. Of course, some of these features might themselves be taken as criteria of good judges. We might, for example, only want to employ judges, human or AI, of a certain threshold of accuracy or interpretability. If AI judges are more accurate decisionmakers than human judges, that might provide a reason to favor them—independent of cost or fairness. We do not pursue these broader arguments here.

Finally, it appears that the human-AI fairness gap was much more strongly driven by people's perception of "hard" factors, such as the accuracy of the decision and the thoroughness of the analysis, than by perceptions of "soft" factors, like the extent to which the decisionmaker understands the defendant's perspective or the extent to which the defendant felt he was heard. These soft factors are presumably those where humans possess inimitable advantages over algorithmic decision makers. But the fact that their contribution is comparatively modest suggests another avenue for the possible narrowing of the human-AI fairness gap. Perceptions of "hard" factors like accuracy and thoroughness will conceivably be updated as technology advances. Especially in domains where a ground truth for the right decision exists and algorithms can be shown to perform better, elimination, even reversal, of the fairness gap seems a real possibility. Proceedings conducted by a robot judge could eventually be considered to be fairer than proceedings in front of a human judge.

Thus, although we document lay evaluation of a human-AI fairness gap, we also find no evidence of an irreducible procedural justice advantage for human. The human-AI fairness gap persists across contexts but it can be narrowed if not erased through algorithmic offsetting. Moreover, the gap is mostly accounted for by "hard" rather than "soft" factors. If the human

advantage over AI is ultimately explained by beliefs about the quality of adjudication rather than anything inherent about the adjudicator, then machines could come to be accepted as procedurally fair decisionmakers, no less so than humans.

Finally, although there are some examples of algorithms acting as decisionmakers, AI tools are often employed in legal settings as aids or adjuncts to human adjudicators. Our study did not directly address the assistive role of AI, focusing instead on the limit case of having robot judges determine people's rights, duties, and obligations. Doing so permits us to re-examine the procedural justice paradigm in the age of machines.

C. Beyond Perceived Fairness: Accuracy, Bias, and Other Factors

Our studies investigated perceived fairness by manipulating three factors: agent (human versus algorithm), hearing (hearing versus no hearing), and interpretability (interpretable versus uninterpretable decision). But there are many other factors that could shape judgments of procedural fairness, and there are certainly many other criteria beyond perceived fairness that should be applied to robot judges.

To some degree, these other considerations can be seen as limitations to our study. For example, studies suggest that some algorithmic processes perpetuate racial bias.[106] This is an extremely important concern, which might outweigh considerations of cost or compliance. Even if algorithmic adjudication is inexpensive and seen to be as fair as human adjudication, we might reasonably reject the use of robot judges on other moral grounds.

At the same time, we should not take these possibilities as decisive arguments against robot judges. In our non-ideal world, the choice is between flawed humans and imperfect machines. Hence, the question that matters is not whether robot judges are biased, but whether they are more or less biased than human judges. As the economist Sendhil Mullainathan puts it:

> Human judges, not just AI judges, can have hidden biases. Indeed, human judges' biases will usually be harder to identify. One can't reliably test human judges, for instance, by asking them to decide the same case twice, once with a white defendant and once with a black defendant.[107]

The degree of racial bias in the judiciary is a controversial and complex topic and outside the scope of this Article. But there is evidence that at least some human judges treat persons of different races differently.[108]

Moreover, it may be more straightforward to address bias in AI judges. According to computer scientist Jon Kleinberg and co-authors, machines offer "far greater" visibility into "the ingredients and motivations of decisions, and

---

[106] *See, e.g.*, Mayson *supra* note 64.

[107] Volokh, *supra* note 17.

[108] *See, e.g.*, David Abrams, Marianne Bertrand & Sendhil Mullainathan, *Do Judges Vary in Their Treatment of Race*, 41 J. LEGAL STUD. 347, 350 (2012).

hence far greater opportunity to ferret out discrimination."[109] It may therefore be that biased algorithms are easier to fix than biased people.[110]

Similar arguments can be made about other factors that are not directly tested in our study. Consider responsiveness, the ability of a judge to respond affectively or appropriately to the parties and their concerns. Perhaps robot judges are on average less responsive than human judges. But there is likely great variation in responsiveness among human judges. "Some [human] judges may be more 'responsive' than others, and others may show more emotion and compassion."[111] Here too, it is far from obvious that AI falling short of some ideal of responsiveness implies that all AI judges are less responsive than all human judges.

One of our suggestions about how to offset the human-AI fairness gap was to generate algorithmic decisions that are more interpretable than human decisions. At the same time, however, it may not be desirable to make decisions entirely interpretable — and hence, predictable — even if doing so were technically feasible. Interpretability might facilitate "gaming" of the system by litigants who manipulate the characteristics of their cases to achieve better outcomes. For instance, a daredevil might paint her car black rather than red if she knew that the robot judge gave heavier fines to drivers of flashier vehicles, perhaps because there is a correlation between the appearance of an individual's vehicle and the speed at which they drive. Strategic behavior like the one described is especially problematic if the variables considered by the algorithm include proxies for the ultimate facts or factors of interest. It is less problematic if the algorithm only takes into account the ultimate facts or factors themselves. Reducing the speed at which the car is driven is not gaming the speed limit law but obeying it! Insofar as algorithms must rely on proxies for what the law ultimately cares about, robot judging may be susceptible to gaming, and strategic opacity might be necessary from a dynamic, all-things-considered perspective.

The Article has thus far focused on the human-AI judge comparison within the context of a single, discrete case. But AI judges might provide other systemic advantages, including some related to legitimacy and fairness. For example, the introduction of robot judges could increase the *total* number of cases adjudicated in a public forum.

Another important aspect of procedural justice debates concerns the growth of mediation and arbitration. Scholars worry that these processes are not fair, and this concern may also be shared by ordinary people. In our study, the consumer arbitration scenario had the lowest procedural justice ratings, regardless of whether the judge was a human or an algorithm. Of course, there are many possible explanations for this obsveration. Our study did not set out to assess lay perceptions of the fairness of arbitration, and future empirical work could more rigorously assess whether people evaluate arbitration as

---

[109] Kleinberg, Ludwig, Mullainathany & Sunstein, *supra* note 68, at 163.

[110] Sendhil Mullainathan, *Biased Algorithms are Easier to Fix than Biased People*, N.Y. TIMES (Dec. 6, 2019), https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html [https://perma.cc/8FQ2-GRS4].

[111] Tania Sourdin, *Judge v Robot? Artificial Intelligence and Judicial Decision-Making*, 41 U. NEW S. WALES L.J. 1114, 1114 (2018).

particularly unfair. Nevertheless, one candidate explanation of the differences across scenarios in our study could be that people tend to see public judicial proceedings as procedurally fairer than private arbitration. If this were true—if distinction between public versus private adjudication has an immense bearing on perceived fairness, then the introduction of robot judges could dispense greater procedural justice in aggregateby allowing more people to have their day in public court.

Before concluding, we note two important limitations of our study. The first is that our conclusions are based on lay judgments of procedural justice. There is a legitimate worry that people may be victims of "false consciousness": They might believe robot judges to be fair even though the truth is otherwise.[112] This worry constitutes a fundamental qualification to a the procedural justice paradigm in legal psychology.[113] While it is worth interrogating the basic assumptions of the field, such an undertaking falls outside the scope of our Article. We acknowledge the possibility of AI being cynically designed to inflate perceptions of fairness, rather than actual fairness. That is, the offset we propose could be employed to manipulate or even deceive the publicOne could imagine extensive hearings that do nothing to change the outcome of machine adjudication, or "faux explanations" of algorithmic decisions that are placatory but untrue.[114]

Second, our analysis is limited to the United States. We recruited a large, nationally representative sample of American adults, so our results reflect popular opinion about robot judges in the United States. It is not obvious that our conclusions would generalize across jurisdictions and cultures:

> It would be easy to state the obvious and repeat that in all justice systems of the world the role of civil justice is to apply the applicable substantive law to the established facts . . . and pronounce fair and accurate judgments. The devil is, as always, in the details. What is the perception of an American judge about his or her social role and function, and does it correspond to the perception of the judge in the People's Republic of China?[115]

Future research could study whether these perceptions of judicial fairness are homogenous or socially and culturally contingent.

CONCLUSION

AI has already assumed the role of judicial assistants and the prospect of robot judges ruling by themselves on some types of cases is no longer unrealistic. At the same time, there are important doctrinal and legal-ethical

---

[112] *See* Robert J. MacCoun, *Voice, Control, and Belonging: The Double-Edged Sword of Procedural Fairness*, 1 ANN. REV. L. & SOC. SCI. 171, 188–193 (2005).
[113] *See* TYLER, *supra* note 11.
[114] Re & Solow-Niederman, *supra* note 13, at 261.
[115] ALAN UZELAC, *Goals of Civil Justice and Civil Procedure in the Contemporary World*, *in* GOALS OF CIVIL JUSTICE AND CIVIL PROCEDURE IN CONTEMPORARY JUDICIAL SYSTEMS 3, 3 (2014).

objections to the introduction of robot judges. This paper has focused on one of the most common and fundamental challenges: Citizens would see robot judges as procedurally unfair, threatening the legitimacy of the judicial system.

The experiments here uncover the truth in this conventional wisdom. Two studies provide evidence of a perceived fairness gap between human and AI adjudicators. Moreover, the same pattern of results is replicated across three distinct contexts: bail, sentencing, and commercial arbitration. We argue, building on existing research on the psychology of procedural justice, that this substantiates an important procedural justice obstacle for robot judging.

At the same time, the study furnishes evidence for a possible solution to this problem, one that may be interesting not only for application designers but also for policymakers and practitioners evaluating the suitability of legal AI solutions. Two other important features — interpretability and hearing — contribute to the perceived fairness of *both* human and AI judges. This raises the possibility of algorithmic offsetting, compensating for the perceived AI-human fairness gap by supplementing an AI-presided proceeding with a hearing, increased interpretability, or perhaps even greater accuracy.

Thus, the results tell a surprising and nuanced story concerning ordinary perceptions of robot judges. People generally favor human judges as procedurally fairer, but they do not perceive human judges as having absolute or irreducible advantages. In fact, in some circumstances, some might prefer to have their day in robot court.