Harvard Journal of Law & Technology
Volume 36, Number 1 Fall 2022

HAVING YOUR DAY IN ROBOT COURT

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

ABSTRACT

Should machines be judges? Some say "no," arguing that citizens would see robot-led legal proceedings as procedurally unfair because the idea of "having your day in court" is thought to refer to having another human adjudicate one's claims. Prior research established that people obey the law in part because they see it as procedurally just. The introduction of "robot judges" powered by artificial intelligence ("AI") could undermine sentiments of justice and legal compliance if citizens intuitively view machine-adjudicated proceedings as 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 — this fairness gap through "algorithmic offsetting." Affording litigants a hearing before an AI judge and enhancing the interpretability of AI decisions reduce the human-AI fairness gap. Moreover, the perceived procedural justice advantage of human over AI adjudication appears to be driven more by beliefs about the accuracy of the outcome and thoroughness of consideration, rather than doubts about whether a party had adequate opportunity to voice their opinions or whether the judge understood the perspective of the litigant.

The results of the experiments can support a common and fundamental objection to robot judges: There is a concerning human-AI fairness gap. Yet, at the same time, the results also indicate that the public may not believe that human judges possess irreducible procedural fairness advantages. In some circumstances, people see a day in a robot court as no less fair than a day in a 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, Jack Leahey, and Jannek Ulm for excellent research assistance.

TABLE OF CONTENTS

I. 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") already plays a role in the U.S. legal system but has thus far primarily served as an aid. For example, algorithms recommend but do not determine criminal sentences in some states.[3] Elsewhere, AI systems could function as primary

[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/Y5A4-YEWC].

[2] *Id.*

[3] *See id.*; *cf.* Frank Fagan & Saul Levmore, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. CAL. 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*

decision-makers 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, challenging value judgments must be made. Perhaps the most critical objection sounds in procedural fairness. Would a judicial proceeding overseen by a robot judge undermine the constitutional 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 procedural justice in this way.[10] Evidence suggests that the perceived fairness of legal processes has far-reaching 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

Autonomous Vehicles, YALE J.L. & TECH. (forthcoming 2023), 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).

[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* State v. Loomis, 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) (cataloging 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).

judges is thus crucial, both for those concerned with legal compliance and 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]

Richard Re and Alicia 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] We present evidence of people's evaluation of robot judges' decisions through a series of original experimental studies involving a large sample of U.S. participants. These vignette experiments vary the decision-maker (human or algorithm), scenario (consumer arbitration, bail, or sentencing), whether there is a hearing, and whether the judge's decision is interpretable.[16]

[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. BEHAV. 134 (2020); Gabriel Lima, Nina Grgić-Hlača & Meeyoung Cha, *Human Perceptions on Moral Responsibility of AI: A Case Study in AI-Assisted Bail Decision-Making*, PROC. CHI CONF. ON HUM. FACTORS COMPUTING SYS. 1, 1 (2021); Bertram Malle, Matthias Scheutz, Thomas Arnold, John Voiklis & Corey Cusimano, *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*, CAMBRIDGE HANDBOOK OF EXPERIMENTAL JURIS. (forthcoming 2024) (manuscript 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 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 are "inherently interpretable" — and

The study makes two significant findings. First, there is a clear human-AI fairness gap: Proceedings conducted by human judges were seen as fairer than those conducted by AI judges. Second, the procedural fairness advantage of human judges seems neither irreducible nor absolute. Remarkably, participants did not evaluate a hearing before an AI judge as meaningless. On the contrary, having the opportunity to speak and be heard increases procedural fairness ratings for both human and AI-adjudicated processes. Our results hint at the possibility of "algorithmic offsetting." That is, the human-AI fairness gap can be offset, partly and perhaps even entirely, by introducing into AI adjudication procedural elements that might be absent from current processes, such as a hearing or an interpretable decision.

Moreover, an exploratory mediation analysis suggests that the human-AI fairness gap is explained by "hard" factors, like the perceived accuracy and thoroughness of the decision-making process, more so than by distinctively human, "soft" factors, like the decision-maker's understanding of the litigant's position or a feeling that the litigant had a voice. 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 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 indicate 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 is unfair. Insofar as human judges set the standard for fairness, our results imply that the procedural justice objection to robot judges may not be decisive. Were robot judges to become more accurate, comprehensive, interpretable, or responsive, their decision-making might even be seen as fairer than that of human judges in some situations.

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. INTEL. 206, 206 (2019). We follow this distinction here.

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

II. 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, human rights scholar Anthony D'Amato mused that computers might replace judges, assuming that "the law has been made completely determinable" and automation would eliminate discretion in judicial decision-making.[18] But 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 in statutory interpretation.[20] Meanwhile, the evaluative nature of many common law concepts means that applying 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] Modern algorithms identify and harness empirical relationships more effectively than their predecessors by leveraging greater computing power and more flexible modeling strategies.[22]

Simple models have already outperformed lawyers in predicting decisions of the U.S. Supreme Court,[23] and more sophisticated models are now boasting impressive accuracy for a diverse range of tribunals.[24]

[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).

[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 Decisionmaking*, 104 COLUM. L. REV. 1150, 1150 (2004).

[24] *See* Daniel Martin Katz, Michael J. Bommarito II & Josh Blackman, *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, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro & Vasileios Lampos, *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,

Their apparent success has excited interest in the possibility of faster, cheaper, and better justice delivered by robot judges.

The role of AI in American criminal law remains very much advisory — legal judgment continues to be delivered by judges sitting in courtrooms.[25] But in the United Kingdom, 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."[26] And while no jurisdiction has to date been bold enough to let an algorithm alone determine a person's guilt or innocence, at least one nation is prepared to let machines resolve some kinds of cases. Estonia is building a system to adjudicate small claims where the amounts in controversy are below €7,000.[27] 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 digitized public services like voting and tax filing.[28]

These developments raise questions about human adjudication's distinctiveness and its future. From a theoretical perspective, adjudication has never been solely about achieving the correct 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."[29] Fuller hence reasoned that "[w]hatever heightens the significance of this participation lifts adjudication towards its optimum expression" and "[w]hatever destroys the meaning of that participation destroys the integrity of adjudication itself."[30] 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 of the rule of law calls for subjects to be afforded the opportunity to interpret the law, relate its abstract demands to their own circumstances, and have their arguments evaluated

Héctor Allende-Cid, Orivaldo Santana & Lívia de Oliveira-Lage, Predicting Brazilian Court Decisions (Apr. 20, 2019) (unpublished manuscript) (on file with author).

[25] *See generally* Bart Custers, *AI in Criminal Law: An Overview of AI Applications in Substantive and Procedural Criminal Law*, 35 LAW & A.I. 205 (2022).

[26] 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/88CT-W4PB].

[27] Niiler, *supra* note 6.

[28] 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].

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

[30] *Id.*

impartially in a neutral forum.[31] These procedural guarantees, Jeremy Waldron argues, are at the heart of ordinary understandings of legality.[32] According to Waldron, "[t]hey 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."[33] 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.[34] 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.[35] 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.[36] Procedures are valued because they allow parties to convey information to the adjudicator.[37] Procedures are also valued because they treat the parties not as objects but as subjects with their own interests and stories.[38] Litigants who believe they have received procedural justice are more likely to recognize the tribunal's authority and accept its determination.[39] While there are several factors

[31] *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).

[32] *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).

[33] *Id.* at 15–16.

[34] 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).

[35] 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).

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

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

[38] Tom R. Tyler, *Psychological Models of the Justice Motive: Antecedents of Distributive and Procedural Justice*, 67 J. PERSONALITY & SOC. PSYCH. 850, 852 (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).

[39] *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

conducive to a sense of fairness,[40] two are especially relevant to AI judgments: voice and justification.

First, people are more inclined to endorse a procedure as fair if they are able to voice their perspective.[41] Voice matters for instrumental and value-expressive reasons. Instrumentally, the chance to advocate a position gives the speaker possible influence over outcomes.[42] Hence, those with a voice may regard a process as fair because their views could shape the decisions made.[43] But they may also regard a process as fair even when their opinions have little hold on the decision-maker.[44] This is because the opportunity to speak acknowledges the parties' agency and their membership in the community.[45] The denial of such an opportunity is especially aggravating in societies and situations where it is expected,[46] and the value-expressive function of voice may sometimes be the most essential one.[47] But for voice to convey respect and inclusion, individuals must also feel heard; they

Court procedures influence its views 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).

[40] *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).

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

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

[43] E. Allan Lind, Ruth Kanfer & P. Christopher Earley, *Voice, Control, and Procedural Justice: Instrumental and Noninstrumental Concerns in Fairness Judgments*, 59 J. PERSONALITY & SOC. PSYCH. 952, 952 (1990).

[44] *Id.*; Tom R. Tyler, Kenneth A. Rasinski & Nancy Spodick, *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, Pascal Langenbach & Lilia Zhurakhovska, *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").

[45] *See* Lind et al., *supra* note 43.

[46] 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, Riël Vermunt & Henk A. M. Wilke, *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 Seven Facility Relocations*, 15 J. ORG. BEHAV. 623, 634 (1994); Patricia Grocke, Federico Rossano & Michael Tomasello, *Young Children Are More Willing to Accept Group Decisions in Which They Have Had a Voice*, 166 J. EXPERIMENTAL CHILD PSYCH. 67, 75 (2018).

[47] Lind et al., *supra* note 43.

must experience their participation as meaningful and not merely a sham.[48]

People also tend to endorse a procedure as fair if decisions are openly justified.[49] By giving reasons, decision-makers reassure the parties that they have "acted on the presented viewpoints in an impartial and unbiased manner."[50] Unsurprisingly, it is often losers who demand justifications for outcomes rather than winners. To satisfy them, the explanations must come across as sincere and adequate.[51] More specific or thorough explanations are also more readily accepted.[52]

Procedures believed to be fair may not actually 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."[53] Giving disputants satisfaction and closure is an essential aspect of any justice system. Tim Wu identifies procedural fairness as an "obvious" advantage that human judges have over their artificial rivals.[54] But is this advantage so obvious? When it comes to having a voice in the process, advances in natural language technology have empowered computers to convert between speech and text, give intelligent replies to questions, summarize documents, and spot contradictions in statements.[55] These advances raise the possibility that parties could one day have their grievances heard by machines in place of human judges.

Yet, the ability to perform these tasks does not mean that machines understand language like humans do. Even if machines could generate

[48] 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).

[49] *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 41, at 683.

[50] Bies, *supra* note 49, at 4; *see also* Daly & Geyer, *supra* note 46, at 627.

[51] *See* Robert J. Bies, Debra L. Shapiro & Larry L. Cummings, *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 46.

[52] *See* Debra L. Shapiro, E. Holly Buttner & Bruce Barry, *Explanations: What Factors Enhance Their Perceived Adequacy*, 58 ORG. BEHAV. & HUM. DECISION PROCESSES 346, 346 (1994) (finding that specificity affects the perceived adequacy of an explanation); 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, and people favor simpler causal explanations over more complex ones).

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

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

[55] *See generally* Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang & Owain Evans, *Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts*, 62 J. A.I. RSCH. 729 (2018).

perfect sentences in Chinese, they do so by learning word frequencies and co-occurrences or by obeying a grammatical logic they have been taught.[56] They do not 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 philosophical 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.[57] 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 of some algorithms from over 75% to a mere 7%.[58] 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, whether machines truly understand humans 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 — C-3PO is an example from popular culture. They are, on an influential theory, social actors.[59] Studies find that people tend to apply rules of social behavior to human-computer interactions despite recognizing the inapplicability of those rules to machines.[60] We are gentler in rating a computer when the evaluation is requested by the computer itself rather than a human third party.[61] We are partial to "silicon sycophants" that flatter us.[62] We even project gender onto machines, heeding the advice of computers represented as male on "masculine" topics and computers represented as female on "feminine"

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

[57] *See generally* Jieyu Lin, Jiajie Zou & Nai Ding, *Using Adversarial Attacks to Reveal the Statistical Bias in Machine Reading Comprehension Models*, PROC. ASSOC. COMPUTATIONAL LINGUISTICS & INT'L JOINT CONF. ON NAT. LANGUAGE PROCESSING 333 (2021) (describing effect of adversarial attack on reading comprehension system, reducing performance from near-human level to mere chance).

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

[59] *See generally* Clifford Nass, Jonathan Steuer & Ellen R. Tauber, *Computers Are Social Actors*, PROC. SIGCHI CONF. ON HUM. FACTORS COMPUTING SYS. 72 (1994) (arguing that individuals' interactions with computers are social).

[60] *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).

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

[62] B.J. Fogg & Clifford Nass, *Silicon Sycophants: The Effects of Computers that Flatter*, 46 INT'L J. HUM.-COMPUT. STUD. 551, 552–53 (1977).

topics.[63] The "computers as social actors" 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.[64]

As for justification, explanations matter in part because they help demonstrate the absence of judicial bias. But suspicions 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,"[65] 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.[66]

D'Amato's qualification is crucial: There is a nagging worry that algorithmic processes perpetuate the same biases that infect humans.[67] Indeed, academics and popular writers have sounded the alarm about algorithms that discriminate.[68] 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 discrimination.[69] One risk of training algorithms on datasets of human decisions is "bias in, bias out."[70] A natural solution is perhaps

[63] *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).

[64] *See generally* Nass et al., *supra* note 59.

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

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

[67] *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).

[68] *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/YH7J-GVZG]; 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].

[69] *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).

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

disclosure. To the extent people are suspicious of the factors and variables machines consider, transparency about inputs might assuage some fears.[71]

Secrecy, however, is not the only misgiving people have about algorithmic judging in general.[72] 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.[73] Clarity about the optimization function and the training data does not guarantee the interpretability of the mechanism or its results.[74] Certainly, one could always furnish the parties with 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.[75] 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, some proponents of interpretable artificial intelligence have recommended exploiting deep neural networks for their accuracy while rendering them explainable through an approximation by decision trees.[76]

[71] *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.

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

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

[74] *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), and Doshi-Velez & Kim, *supra* note 16, at 9.

[75] *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).

[76] *See* Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, *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 neural networks and decision trees).

Still, amidst the disquiet about AI, it must be kept in mind that humans are not always open and honest in their reasoning either.[77] According to computer scientist Jon Kleinberg and coauthors, algorithms offer "far greater" visibility into "the ingredients and motivations of decisions, and hence far greater opportunity to ferret out discrimination."[78] "[T]here is instead every reason to think," as Aziz Huq writes, that "[human] judicial discretion has had dismaying and socially destructive effects."[79]

III. Two Experimental Studies

While there is rich theoretical literature on AI judging and legal processes, many 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 worry 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 that offer fresh empirical evidence bearing on each of these central and largely untested questions. All study materials, including pre-registrations, vignettes, and data have been made available online.[80]

[77] *See* Kleinberg et al., *supra* note 71, 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.").

[78] Kleinberg et al., *supra* note 71, at 163.

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

[80] *Having Your Day in Robot Court*, Open Sci. Framework, https://osf.io/cw2m4 [https://perma.cc/5M2J-VVL2].

A. Study 1

We investigate how people perceive the fairness of human as opposed to AI judges in three different adjudicatory contexts: consumer refunds for a damaged product, pretrial bail determination for criminal offenses, and custodial sentencing post-conviction. These scenarios were presented to experimental subjects in vignettes, all featuring the same protagonist, John Smith. The vignettes are reproduced in the following Section.

1. Experimental Scenarios

Subjects were randomly assigned to one of three scenarios. In every scenario, the decision would ultimately go against John: He would not obtain a refund for an allegedly damaged camera, he would be denied bail pending trial for possession of a controlled substance with intent to distribute, and he would receive the maximum possible sentence for manslaughter.

Consumer Refund. This scenario recounted the arbitration of 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." 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.

Bail. This scenario concerned a pretrial bail decision following 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.

Sentencing. The last scenario revolved around a sentencing decision in a manslaughter case:

> 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.

2. Experimental Treatments

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

Type of Decision-Maker. We were primarily interested in how ordinary people assessed the fairness of a decision based on the human or algorithmic nature of the decision-maker. Therefore, both kinds of decision-makers were introduced as highly competent at their adjudicative tasks. For example, in the pretrial bail scenario, subjects randomized to the "Human" condition read that:

> In 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:

> In 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. The decision could have been made exclusively based on the record or could have included a hearing. For example, in the consumer refund case, subjects randomized to the "Algorithm" and "Hearing" conditions were informed that:

> Before 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 read that:

> Before 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 refers to transparency into — and knowledge of — how the outcome is derived, not the provision of a reason for the outcome. Thus, under the "Interpretable" condition, the vignette concluded by stating that:

> While 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 [sic] is satisfied. It would therefore be possible for John, or anyone else, to figure out how much each factor mattered to the

> [decision-maker]'s ultimate decision. Moreover, it would be possible for someone else to replicate the [decision-maker]'s reasoning to see how a change in any of his factors impacts the sentencing decision.

In contrast, under the "Not Interpretable" condition, the vignette ended by admitting that:

> While the [decision-maker]'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 [decision-maker] 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 solely based 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 AI. 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 heard and given a fair shake. Moreover, people are not accustomed to having algorithms resolve their disputes, and computers — unlike humans — are vulnerable to hardware malfunction or programming bugs. There could therefore be some uneasiness about having algorithms determine matters of great importance.[81] On the other hand, people sometimes trust the advice of computers, believing them to be better at objective tasks than even human experts.[82] Perhaps incorrectly, people also tend to conceive of algorithms as being rule-bound and, hence, less capricious than humans, who may succumb to passions or preconceptions.[83]

[81] *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).

[82] *See* Noah Castelo, Maarten W. Bos & Donald R. Lehmann, *Task-Dependent Algorithm Aversion*, 56 J. MKTG. RSCH. 809, 818 (2019) ("[E]mphasizing the quantitative approach to accomplishing the tasks succeeded at increasing trust in algorithms."); *see also* Jennifer M. Logg, Julia A. Minson & Don A. Moore, *Algorithm Appreciation: People Prefer Algorithmic to Human Judgment*, 151 ORG. BEHAV. & HUM. DECISION PROCESSES 90, 93 (2019).

[83] *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, some also view that in an unfavorable light).

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 relate to the observable classification of a physical object, whereas the latter requires predicting 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." Thus, the law calls for a normative balancing of several factors — considerations that bear on recidivism and rehabilitation but also speak to blame and culpability. In sum, it is plausible that AI 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.

Finally, the contribution of voice and interpretability to procedural justice may depend on the cognitive and emotional capacities of the decision-maker. The opportunity to speak and be heard might only be regarded as meaningful if the adjudicator can parse language and genuinely understand and empathize with the parties. Humans, unlike algorithms, possess these capabilities. 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 AI decisions. At the same time, people ascribe mental states to computers, projecting norms, beliefs, and stereotypes onto them.[84] 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," "Bail," or "Sentencing" scenarios, the "Human" or "Algorithm" condition, the "Hearing" or "No Hearing" condition, and the "Interpretable" or "Not Interpretable" condition. This first study presented features a between-subject, 3×2×2×2 factorial design, wherein each participant read a

[84] *See supra* notes 59–63 and accompanying text.

single vignette describing a randomly selected scenario featuring randomly varied factors. For example, a participant might be randomly assigned to the consumer refund scenario with a hearing and an uninterpretable algorithmic decision.

Table 1: Four Factors in the 3×2×2×2 Between-Subjects Design

Scenario	Type of Decision-Maker	Interpretability	Hearing
Consumer Refund	Human	Interpretable	Hearing
Bail			
Sentencing	Algorithm	Not Interpretable	No Hearing

Before reading any of the scenarios, subjects were first asked 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 decision-maker's procedure for arriving at the decision was fair, whether the decision-maker considered all relevant facts in making the decision, and whether the decision-maker 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 decision-maker 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 conducted on 1,710 subjects in September 2020. Subjects were recruited through Lucid Theorem and sampled to be nationally representative of the U.S. population.[85] 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 decision-maker. 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 over a 1-to-7 scale ($p<0.001$, two-sided t-test; $p=0.002$, two-sided Wilcoxon rank sum test) than those in the "Human" baseline.[86]

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).[87] Compared to the "Not Interpretable" baseline, subjects in the "Interpretable" condition gave fairness ratings that were 0.305 higher on average ($p=0.002$, two-sided t-test; $p<0.001$, two-sided Wilcoxon rank sum test).[88]

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:

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

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

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

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

[87] 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.

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

interaction term $I \times T$ — captures moderation in the treatment effect.[89] We consider the following models of this form.

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.[90]

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.[91]

A linear regression of procedural fairness ratings on Scenario indicators, a Type of Decision-Maker indicator, and Type of Decision-Maker-Scenario interactions also returns statistically insignificant coefficients for the Algorithm-Bail (-0.053, *p*=0.827) and Algorithm-Sentencing (-0.203, *p*=0.380) interaction terms.[92]

[89] *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 2K Factorial Designs*, 112 STAT. & PROBABILITY LETTERS 72, 75–76 (2016).

[90] 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 the procedural fairness rating, 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 "Consumer Refund" ($I_{Bail} = 0$ and $I_{Sentencing} = 0$) and "No Hearing" ($T_{Hearing} = 0$).

[91] The estimated model is $y = \beta_0 + \beta_1 I_{Bail} + \beta_2 I_{Sentencing} + \beta_3 T_{Interpretability} + \beta_4(T_{Interpretability} \times I_{Bail}) + \beta_5(T_{Interpretability} \times I_{Sentencing})$ where y is the procedural fairness rating, I_{Bail} is an indicator variable for the bail scenario, $I_{Sentencing}$ is an indicator variable for the sentencing scenario, and $T_{Interpretability}$ is an indicator variable for the interpretability of the decision. Note that the reference levels are "Consumer Refund" ($I_{Bail} = 0$ and $I_{Sentencing} = 0$) and "Not Interpretable" ($T_{Interpretability} = 0$).

[92] 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 the procedural fairness rating, 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 decision-maker. Note that reference levels are "Consumer Refund" ($I_{Bail} = 0$ and $I_{Sentencing} = 0$) and "Human" ($T_{Algorithm} = 0$).

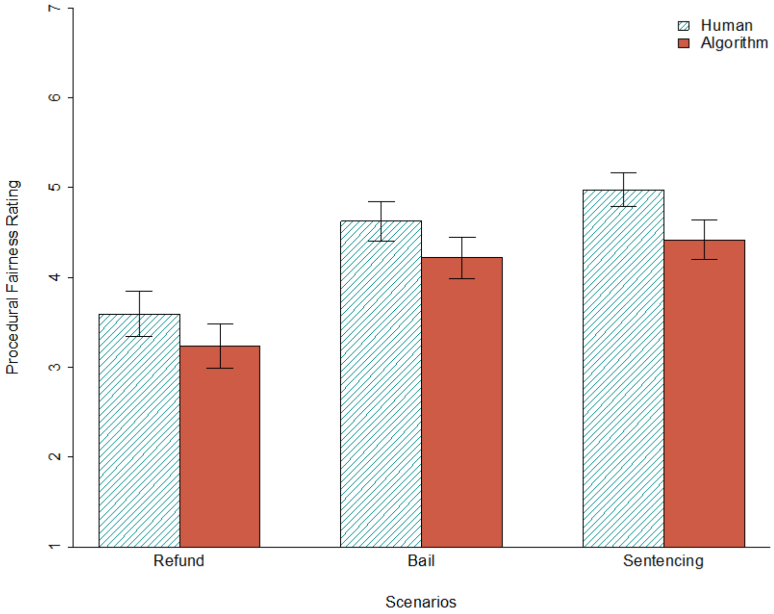

Figure 1: The Human-AI Fairness Gap: Average Procedural Fairness Rating in Study 1 by Scenario and Decision-Maker

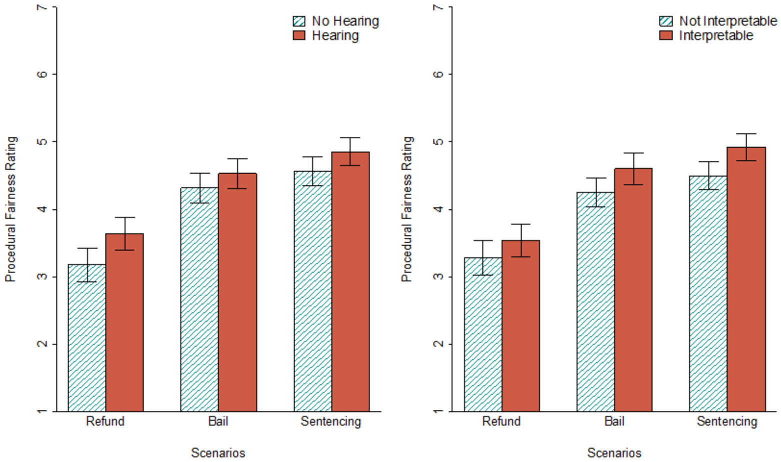

Figure 2: Hearing and Interpretability Increase Fairness: Average Procedural Fairness Rating in Study 1 by Scenario and Hearing/Interpretability

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 Decision-Maker, and the interaction between the variables (indicated with "-"). Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses.

	Hearing by Scenario	Interpretable by Scenario	Decision-Maker by Scenario
Constant	3.1753^{***} (0.1279)	3.2780^{***} (0.1286)	3.5920^{***} (0.1271)
Hearing	0.4632^{**} (0.1775)		
Interpretable		0.2516 (0.1783)	
Algorithmic Decision-Maker			-0.3545^{*} (0.1779)
Bail	1.1414^{***} (0.1719)	0.9746^{***} (0.1685)	1.0344^{***} (0.1683)
Sentencing	1.3880^{***} (0.1659)	1.2170^{***} (0.1661)	1.3815^{***} (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 Decision-Maker - Bail			-0.0525 (0.2400)
Algorithmic Decision-Maker - 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
Key:	$^{*}p < 0.05$ $^{**}p < 0.01$ $^{***}p < 0.001$		

We also investigate whether the influence of a hearing or the interpretability of the decision on procedural justice judgments varies by the type of decision-maker. To do so, we linearly regress procedural fairness ratings on a Type of Decision-Maker indicator, a Hearing indicator, and the interaction between both indicators (Table 3).[93] The estimate for the 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 Type of Decision-Maker indicator, an Interpretability indicator, and the interaction between both indicators (Table 3).[94] The estimate for the coefficient of the interaction term is positive but statistically insignificant (0.243, *p*=0.207).

To summarize, consistent with the prior literature on human decision-makers,[95] 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 decision-maker matters. A decision made by an algorithm is viewed as less procedurally fair than a decision made by a human. The data hints that a hearing is more important than interpretability for perceived fairness when the decision-maker is a human as opposed to when an algorithm makes the decision. But these differences are statistically insignificant.

[93] 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 decision-maker, and $T_{Hearing}$ is an indicator variable for hearing.

[94] The estimated model is $y = \beta_0 + \beta_1 T_{Algorithm} + \beta_2 T_{Interpretability} + \beta_3 (T_{Algorithm} \times T_{Interpretability})$ where y is procedural fairness ratings, $T_{Algorithm}$ is an indicator variable for algorithmic decision-maker, and $T_{Interpretability}$ is an indicator variable for the interpretability of the decision.

[95] *See supra* notes 10–11, 34–44.

Table 3: Estimated coefficients for Study 1 from an ordinary least squares regression of procedural fairness rating on an indicator variable for Type of Decision-Maker, an indicator variable for Hearing/Interpretability, and the interaction of both variables (indicated with "-"). Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses.

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

B. Study 2

1. Scenario, Experimental Treatments, and Hypotheses

To probe for interactions between the Hearing and Interpretability factors and Type of Decision-Maker, we replicate the first study but this time limit the scenario to Bail only and employ a larger sample size. The pretrial bail scenario was chosen because it involved moderate stakes and an evaluative task not overly dependent on normative determinations. Simulations based on data from the first study indicated that an experiment conducted on 5,000 subjects would have 80% power to detect the interaction between Hearing and Type of Decision-Maker and 75% power to detect the interaction between Interpretability and

Type of Decision-Maker. The experimental treatments remained the same. The second study thus features a 2×2×2 factorial design: the Type of Decision-Maker may be a "Human" or "Algorithm," there may be a "Hearing" or "No Hearing," and the decision may be "Interpretable" or "Not Interpretable."

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

Type of Decision-Maker	**Interpretability**	**Hearing**
Human	Interpretable	Hearing
Algorithm	Not Interpretable	No Hearing

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, 5,086 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 decision-maker. Moreover, 86.6% and 87.5% of subjects randomly assigned to the "Interpretable" and "Not Interpretable" conditions correctly recalled how the decision was made.

Confirming the results of the first study, both 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).[96] 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).[97]

[96] The Neyman estimator gives a conservative standard error of 0.051.

[97] The Neyman estimator gives a conservative standard error of 0.051.

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.[98]

We find no evidence of an interaction between Hearing or Interpretability and Type of Decision-Maker. A hearing increased fairness ratings for both human and algorithmic decision-makers, and there is no discernable difference in effect across the two conditions (Table 5). Likewise, interpretability boosted subjective judgments of procedural justice for both human and algorithmic decision-makers with no discernable difference in effect (Table 5).

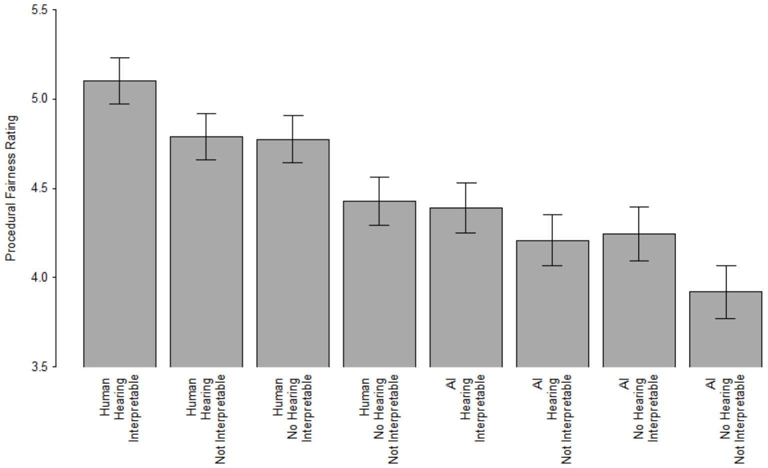

Figure 3: Average Procedural Fairness Rating in Study 2 by Decision-maker, Hearing, and Interpretability. Note: To facilitate comparison of ratings across cells, the Figure's y-axis begins from 3.5 and ends at 5.5. The experimental question presented subjects with a 1-to-7 scale.

[98] The Neyman estimator gives a conservative standard error of 0.050.

Table 5: Estimated coefficients for Study 2 from an ordinary least squares regression of procedural fairness rating on an indicator variable for Type of Decision-Maker, an indicator variable for Hearing/Interpretability, and the interaction of both variables (indicated with "-"). Robust standard errors are computed using the HC2 sandwich estimator and reported in parentheses where applicable.

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

3. Accounting for the Perceived Fairness Gap Between Human and Algorithmic Decision-Makers

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 compared to algorithmic judging, even when there is no hearing. Relatedly, a human may be perceived as capable of empathizing with the defendant's situation in ways an algorithm cannot. Alternatively, people may believe human judging to be more thorough or accurate than algorithmic judging.

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 decision-maker understood John's perspective in making the decision, and that the decision-maker 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 it represents the subject's estimate of the percentage of correct decisions rendered by the decision-maker.

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 Type of Decision-Maker on judgments of procedural fairness, then subjects who share the same estimate of the accuracy of the decision-maker will rate the bail proceedings as equally fair whether it is administered by a human or an algorithm. To the extent that a proceeding before an algorithmic decision-maker is rated as less procedurally fair than one conducted by a human, it is only 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.[99]

Randomization of treatment alone is insufficient for identifying and estimating average direct and average causal mediation effects.

[99] 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:

$$Y_i\big(1, M_i(1)\big) - Y_i\big(0, M_i(0)\big) = Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big) + Y_i\big(1 - t, M_i(1)\big) - Y_i\big(1 - t, M_i(0)\big)$$

where $\xi_i(t) = Y_i\big(1, M_i(t)\big) - Y_i\big(0, M_i(t)\big)$ is defined as the direct effect and $\delta_i(t) = Y_i\big(t, M_i(1)\big) - Y_i\big(t, M_i(0)\big)$ 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, Luke Keele & Teppei Yamamoto, *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 $\delta_i(t) = Y_i\left(t, M_i\big(1, W_i(1)\big), W_i(t)\right) - Y_i\left(t, M_i\big(0, W_i(0)\big), W_i(t)\right)$. Kosuke Imai & Teppei Yamamoto, *Identification and Sensitivity Analysis for Multiple Causal Mechanisms: Revisiting Evidence from Framing Experiments*, 21 POL. ANALYSIS 141, 147–49 (2013).

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 decision-maker or the scenario where an algorithm is a decision-maker. We would also have to manipulate subjects' beliefs about whether John felt he had a good opportunity to voice his perspective. And we would have to do so very precisely — subjects' beliefs would have to be either the beliefs they would have had were the decision-maker a human or the beliefs they would have had were the decision-maker an algorithm. Absent a way to reliably construct such beliefs, 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 decision-maker and pre-existing attributes of the subjects.[100] 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 minimal trust in legal authorities are likely to view the adjudicative process as unfair; they are also likely to believe that the judge — human or algorithm — failed to consider all the facts or failed to understand the perspective of the defendant. We therefore adjust for this attribute in the mediation analysis. Finally, we do not take causal independence between the putative mediators for granted. Subjects who say that the decision-maker understood John's perspective might, for that reason, also believe that the decision-maker considered all the facts in arriving at the outcome. We make the necessary further assumption for causal mediation effects to be point-identified.[101]

Average causal mediation effects are computed for Voice, Understanding, Thoroughness, and Accuracy using the "mediation" package for R.[102] A varying coefficient linear structural equations model is 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 came from having an algorithm rather than a human decide on bail is mediated by Voice. The contributions of

[100] Random assignment ensures that assignment to the experimental conditions is statistically independent of potential outcomes and candidate mediators. *See* Imai & Yamamoto, *supra* note 99, at 146–47.

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

[102] *Id.* at 158; *see also* Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele & Kosuke Imai, *Mediation: R Package for Causal Mediation Analysis*, 59 J. STAT. SOFTWARE 1, 26 (2014).

Understanding, Thoroughness, and Accuracy are 12.0%, 27.3%, and 29.3%, respectively.[103]

Table 6: Estimates from varying coefficient linear structural equations model with procedural fairness ratings as the outcome variable, Type of Decision-Maker 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.

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)
Total Effect	-0.594 (-0.685, -0.50)			

IV. Implications

This Part discusses the implications of the experimental results. The first is a challenge for advocates of robot judges. Our studies reveal that people generally see robot judges as less procedurally fair than human judges across different scenarios. Although others have raised objections to AI judicial decision-making on the grounds of procedural justice, our studies provide 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 rigorous 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 obstacles. They also support an objection grounded in compliance. Research in legal psychology suggests that the legitimacy of the judicial system suffers if people see proceedings as unfair.[104]

At the same time, the empirical results reveal a possible — and unintuitive — approach for making robot judging more acceptable to

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

[104] *See* Tyler, *supra* note 11, at 3–4.

disputants. The studies find that lay perceptions of procedural fairness are also affected by the availability of a hearing and by the interpretability of the decision. Significantly, these factors increase the perceived fairness of *both* human and AI judges. Indeed, we find that adding a hearing does not increase the perceived fairness of human-led proceedings more than it does for AI-led proceedings. Simply, we do not 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 its very nature, more familiar or tractable than machine adjudication.

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

Our findings imply that the perceived human-AI fairness gap could be closed through algorithmic offsetting, that is, by incorporating traditional elements of procedural justice into an AI-led 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.[105] 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 being 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 Section IV.C we address some difficulties and limitations of the studies. Other factors may affect people's evaluation of the fairness of judges (e.g., bias or accuracy), but as evinced by our studies, these factors do not necessarily represent advantages unique to human

[105] *See generally* Goldberg v. Kelly, 397 U.S. 254 (1970) (considering the circumstances when the Due Process Clause of the Fourteenth Amendment requires an evidentiary hearing).

judges. Moreover, many qualities of human judges will vary from judge to judge; for example, human judges may be implicitly biased, 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."[106] Automated processes are already deployed in U.S. administrative practice,[107] and robot judges may soon become a reality in other jurisdictions.[108]

Whether robot judges can gain public acceptance, however, is contested. Of course, we cannot expect robot judges to be perfectly fair, because even human judges may fall short of such an ideal. 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."[109]

Our vignette experiments manipulated the type of decision-maker (human or algorithm) to assess whether Americans see AI-led proceedings as more unfair than human-led proceedings. We find a perceived fairness gap; human judges were seen as fairer than AI judges.[110] Moreover, this gap arose consistently across three distinct scenarios: consumer refund arbitration, bail determination, and sentencing.

This discovery raises critical challenges for advocates of robot judges and the 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. Consider, for example, the fact 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 more opportunities for hearings.[111] By the same token, the existence of the human-AI fairness gap could be one reason — and not necessarily a decisive one — for adjudicative proceedings to employ human rather than robot judges.

[106] Liptak, *supra* note 1.

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

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

[109] Volokh, *supra* note 17.

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

[111] Of course, this reason is not conclusive and might be outweighed by others; for example, it would be prohibitively expensive for all adjudicative proceedings to include hearings.

Beyond this ethical argument, the results also substantiate a legal compliance concern about robot judging. Legal psychologists have demonstrated a relationship between perceived fairness and legal compliance.[112] If people regard robot judges as less fair, they may be less inclined to follow the laws that robot judges administer. Introducing robot judges to reduce judicial administrative costs might come at a price of increased non-compliance.

The human-AI fairness gap — a principal finding of our studies — thus poses both due process and legal compliance difficulties for robot judges. At the same time, other results from our studies tell against the thesis that human judges have distinctive or absolute fairness advantages over robot judges.[113] Therefore, it might be possible to mitigate the human-AI fairness gap by affording greater procedural justice safeguards in AI-led proceedings. The next Section develops this idea.

B. Offsetting the Human-AI Fairness Gap

Besides the effect of the type of decision-maker on perceptions of fairness, the experiment uncovered several other effects. Both the interpretability of the decision and the opportunity for a hearing improved judgment of procedural fairness. The tenor of our results is consistent with earlier research on procedural justice conducted with human decision-makers; more interpretable decisions were seen as fairer, and adding a hearing increased the perceived fairness of the proceeding. One striking and novel finding of our studies is that these same effects were observed for robot judges. That is, a hearing before a robot judge increased the perceived fairness of the robot judge's decision, and more interpretable algorithmic decisions were also seen as fairer. Moreover, the effect on perceived fairness of adding a hearing was not appreciably larger for human judges than for robot judges.

It is also striking that the human-AI fairness gap was consistent across scenarios. In the consumer refund arbitration scenario, $2,500 were at stake; in the sentencing scenario, ten years in prison were at stake. The dispositive issue in the consumer refund 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 they exist, should have strengthened as the stakes increased and the issue in question went beyond mere factfinding, but we find no evidence of this pattern.

[112] *See* TYLER, *supra* note 11, at 3–4.

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

These results complement our main finding concerning the human-AI fairness gap. Although human judges were seen as fairer than robot judges, participants also evaluated those decision-makers in an unexpectedly similar way. That is, the perceived procedural fairness benefit of a hearing or an interpretable decision was 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. In our second study, a human-led process with no hearing and resulting in an uninterpretable sentence was seen as no fairer than an AI-led process with the opportunity for a hearing and an interpretable sentencing decision.[114]

Our empirical findings open the door to algorithmic offsetting. Algorithmic 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 studies, 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, be they human or algorithm, with a certain threshold of accuracy. If AI judges are more accurate decision-makers than human judges, that is a reason to favor them independent of cost or other fairness concerns.

Finally, it appears that the human-AI fairness gap was much more strongly driven by perceptions 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 decision-maker understood the litigant's perspective or the extent to which the defendant felt he was heard. These soft factors are presumably those where humans are more likely to possess inimitable advantages over algorithmic decision-makers, but their comparatively modest contribution to judgments of procedural justice illuminates another approach to narrowing the human-AI fairness gap. Public perceptions of hard factors like accuracy and thoroughness will conceivably evolve 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 to be 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 a human-AI fairness gap, we also find no evidence of an irreducible procedural justice advantage for human judges. The human-AI fairness gap persists across contexts, but it can be narrowed if not erased through algorithmic offsetting. Moreover, the

[114] *See supra* Figure 3.

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 an inherent quality of the adjudicator, then machines could come to be accepted as procedurally fair decision-makers, no less so than humans.

Finally, although there are some examples of algorithms acting as decision-makers today, AI tools are often employed in legal settings as aids or adjuncts to human adjudicators. Our studies did not address the assistive role of AI, restricting attention to 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

We investigated perceived fairness by manipulating three factors: Type of Decision-Maker (human or algorithm), Hearing (hearing or no hearing), and Interpretability (interpretable or not interpretable decision). But there are many other factors shaping 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 taken as limitations to our studies. For example, some algorithmic processes perpetuate racial bias.[115] This is a grave concern that might outweigh issues of cost or compliance. Even if algorithmic adjudication were 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 worries as decisive arguments against robot judges. After all, 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 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.[116]

[115] *See, e.g.*, Mayson *supra* note 67.

[116] Volokh, *supra* note 17.

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.[117]

Moreover, it could be more straightforward to address bias in robot judges. 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."[118] It may therefore be that biased algorithms are easier to fix than biased people.[119]

Similar arguments can be made about other factors that are omitted from our experiments. Consider responsiveness, the ability of a judge to respond effectively 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."[120] Here too, it does not follow from AI falling short of some ideal of responsiveness that all AI judges are less responsive than all human judges.

One of our proposals for offsetting 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 exploit algorithmic decision-making to achieve better outcomes. For instance, a daredevil might paint her car black instead of 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 one drives 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 vulnerable to gaming, and deliberate opacity might be necessary from a dynamic perspective.

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

[118] Kleinberg et al., *supra* note 71, at 163.

[119] 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/LKZ7-4VCZ].

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

The Article has thus far focused on comparing human and AI judges 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, improving perceptions that justice is indeed being served.

Another important strand of procedural justice debates touches on the rise of mediation and arbitration. Scholars have doubted the fairness of these dispute-resolution mechanisms, and ordinary people may also share this distrust. In our first study, the consumer refund 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 observation. That study did not set out to assess lay perceptions of the fairness of arbitration, and future research could more rigorously assess whether people see arbitration itself as particularly unfair. Nevertheless, one 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 the distinction between public and private adjudication bears on perceived fairness — then the introduction of robot judges could bring about greater procedural justice in aggregate by allowing more people to have their day in public court.

Before concluding, we note two important caveats. 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 the opposite.[121] This worry constitutes a fundamental qualification to the procedural justice paradigm in legal psychology.[122] 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 at the expense of actual fairness. That is, the offset we propose might be employed to manipulate or even deceive the public. One 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.[123]

Second, our analysis is limited to the United States. We recruited a large, nationally representative sample of American adults, so our results reflect popular opinions about robot judges in the United States.

[121] *See* Robert J. MacCoun, *Voice, Control, and Belonging: The Double-Edged Sword of Procedural Fairness*, 1 ANN. REV. L. & SOC. SCI. 171, 188–93 (2005).

[122] *Id.*; *see* TYLER, *supra* note 11.

[123] Re & Solow-Niederman, *supra* note 13, at 261.

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?[124]

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

V. Conclusion

AI systems already provide judicial assistance, and the prospect of robot judges issuing rulings on their own for some types of cases is no longer unrealistic. At the same time, there are important doctrinal and legal-ethical objections to robot judges. This Article has focused on one of the most common and fundamental challenges to introducing robot judges: Citizens might see robot judges as procedurally unfair, to the point of threatening the legitimacy of the judicial system.

Our experiments lend some credence to this conventional wisdom. Two studies uncover a perceived fairness gap between human and AI judges. Moreover, the same pattern of results is replicated across three distinct contexts: consumer retail arbitration, bail determination, and criminal sentencing. Building on existing research on the psychology of procedural justice, we argue that this finding substantiates an important procedural justice objection to AI-led proceedings.

At the same time, our studies furnish evidence for a possible solution to this problem, which can inform system designers, policymakers, and practitioners in evaluating the suitability of AI legal solutions. The availability of a hearing and the interpretability of the decision enhance the perceived fairness of both human and AI adjudication. This phenomenon raises the possibility of algorithmic offsetting, that is, the narrowing of the perceived human-AI fairness gap by supplementing an AI-led proceeding with the opportunity for a hearing and more interpretable decisions.

[124] 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).

Overall, our studies uncover a surprising and nuanced account concerning robot judges. People generally perceive human judging as procedurally fairer, but the human advantage is neither irreducible nor absolute. In fact, in some circumstances, people might prefer to have their day in robot court.