Journal of Criminal Law and Criminology

Volume 109
Issue 2 *Spring*

Article 1

Spring 2019

Artificial Intelligence and Role-Reversible Judgment

Kiel Brennan-Marquez

Stephen Henderson

Follow this and additional works at: https://scholarlycommons.law.northwestern.edu/jclc

Part of the Criminal Law Commons, Criminal Procedure Commons, Judges Commons, and the Litigation Commons

Recommended Citation

Kiel Brennan-Marquez and Stephen Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. Crim. L. & Criminology 137 (1019).
https://scholarlycommons.law.northwestern.edu/jclc/vol109/iss2/1

This Article is brought to you for free and open access by Northwestern Pritzker School of Law Scholarly Commons. It has been accepted for inclusion in Journal of Criminal Law and Criminology by an authorized editor of Northwestern Pritzker School of Law Scholarly Commons.

0091-4169/19/10902-0137
THE JOURNAL OF CRIMINAL LAW & CRIMINOLOGY Vol. 109, No. 2
Copyright © 2019 by Kiel Brennan-Marquez & Stephen E. Henderson *Printed in U.S.A.*

CRIMINAL LAW

ARTIFICIAL INTELLIGENCE AND ROLE-REVERSIBLE JUDGMENT

KIEL BRENNAN-MARQUEZ & STEPHEN E. HENDERSON[†]

Intelligent machines increasingly outperform human experts, raising the question of when (and why) humans should remain 'in the loop' of decision-making. One common answer focuses on outcomes: relying on intuition and experience, humans are capable of identifying interpretive errors—sometimes disastrous errors—that elude machines. Though plausible today, this argument will wear thin as technology evolves.

In this Article, we seek out sturdier ground: a defense of human judgment that focuses on the normative integrity of decision-making. Specifically, we propose an account of democratic equality as 'role-reversibility.' In a democracy, those tasked with making decisions should be susceptible, reciprocally, to the impact of decisions; there ought to be a meaningful sense in which the participants' roles in the decisional process could always be inverted. Role-reversibility infuses the act of judgment with a 'there but for the grace of god' dynamic and, in doing so, casts judgment as the result of self-rule.

After defending role-reversibility in concept, we show how it bears out in the paradigm case of criminal jury trials. Although it was not the historical impetus behind the jury trial—at least, not in any strong sense—

[†] Kiel Brennan-Marquez is an Associate Professor of Law and William T. Golden Scholar at the University of Connecticut; Stephen E. Henderson is the Judge Haskell A. Holloman Professor of Law at the University of Oklahoma. They wish to thank the readers who improved this article and the many who engaged with conversations regarding its themes, including Mark Blitz, Andy Coan, Brian Choi, Aloni Cohen, Andrew Ferguson, David Gray, Roger Ford, Paul Kahn, Karen Levy, Jonathan Manes, Paul Ohm, Frank Pasquale, Richard Re, Andrew Selbst, Ric Simmons, Christopher Slobogin, Kelly Sorensen, Katherine Strandburg, Daniel Susser, Mark Verstraete, Tal Zarsky, and all the participants at the Privacy Law Scholars Conference and at roundtables at Fordham, Penn State Law, and the Arizona James E. Rogers College of Law.

we argue that role-reversibility explains some of the institution's core features and stands among the best reasons for its preservation. Finally, for the sci-fi enthusiasts among us, role-reversibility offers a prescription as to when the legal system will be ready for robo-jurors and robo-judges: when it incorporates robo-defendants.

TABLE OF CONTENTS

The madman is not only a beggar who thinks he is a king, but also a king who thinks he is a king.

- Jacques Lacan

INTRODUCTION

Imagine it is 2049, and machines have assumed responsibility for large swaths of the legal system. Compared to human decision-makers of old, machines are both faster and more accurate, the crowning achievement of a decades-long effort to rid the system of bias and unpredictability.[1] Competing notions of 'accuracy' still exist, of course—people continue to disagree about the purposes, and attendant priorities, of different areas of law—but machines have been trained to account for such disagreement. It

[1] As physicist Max Tegmark has described the goal:

> What are the first associations that come to your mind when you think about the court system in your country? If it's lengthy delays, high costs and occasional injustice, then you're not alone. Wouldn't it be wonderful if your first thoughts were instead "efficiency" and "fairness"? Since the legal process can be abstractly viewed as a computation, inputting information about evidence and laws and outputting a decision, some scholars dream of fully automating it with *robojudges*: AI systems that tirelessly apply the same high legal standards to every judgment without succumbing to human errors such as bias, fatigue or lack of the latest knowledge.

MAX TEGMARK, LIFE 3.0 105 (2017).

turns out, in fact, that machines are better equipped to deal with the reality of human pluralism than humans themselves: standing outside the fray, machines readily synthesize different normative viewpoints. Furthermore, machines are impeccably consistent. The 'like cases should be treated alike' ideal, forever precarious in a world of decentralized human judging, has been vindicated at last.[2] Using hyper-complex modeling techniques, machine decision-making effectively guarantees that cases with meaningfully identical features always come out the same way.

What—if anything—is wrong with this picture? As a practical matter, there are plenty of reasons to be skeptical of the premises built into the hypothetical.[3] But for argument's sake, suppose the premises hold; suppose machines really will be capable of more accurate and consistent decisions. Is there something wrong, even so, with entrusting certain decisions to machines? Should humans remain 'in the loop' of some decision-making

---

[2] Sir Patrick Devlin described the 'consistency' ideal, and the difficulty of achieving it, in his famous 1956 jury lectures:

> Is there then more than one sort of justice? Yes: the justice of the case is the best compromise that can be obtained between the demands of the law and judgment on the merits. Those two points are always separated by some distance, long or short, and justice can be brought to rest anywhere between them . . . . Judgments according to the merits do not necessarily make justice. They would do so if there were only one judge and he always remembered to decide everything the same way and he went on living for ever—in short, if justice on earth were divine and not human. But for human justice the law is indispensable. . . . It is an essential attribute of justice in a community that similar decisions should be given in similar cases. But that can be done only by following the law. . . . So the law gives the general rule. No general rule ever fits exactly any particular case.

SIR PATRICK DEVLIN, TRIAL BY JURY 151–52, 153 (1956).

We should note that Ben Johnson and Richard Jordan have provocatively argued against the 'like cases' ideal doing real (good) work. *See* Ben Johnson & Richard Jordan, *Should Like Cases Be Decided Alike?: A Formal Analysis of Four Theories of Justice* (forthcoming) (draft available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3127737 [https://perma.cc/94B3-N9ZG]). We remain comfortable, however, with our rather minimal claim of seeking consistency in the applications of our rules of criminal justice.

[3] For example, it will be far harder to know in criminal justice than, say, in pattern or biometric recognition—two current areas of development—that a machine is achieving 'right results.' Even if the machine closely predicts past criminal justice outcomes, there is ample reason to suspect the legitimacy of those past outcomes, both factually and legally, yet we have no better data on which a system can learn. Historic truth is alone difficult to discern—especially in the criminal realm where we try to peer into a defendant's understandings and motivations—but criminal justice outcomes add in principles of system-correction and fairness that are anti-accuracy in the individual case, including suppression of relevant evidence. The same goes for sentencing, where it is hard to imagine we have historically well-predicted desert or dangerousness. But, again, our project is to take the science as a best-case scenario.

even if it fails to increase—and may well diminish—accuracy and consistency?

We think so, and we suspect this intuition is widespread. The question is whether it can be rationalized. In this Article, after briefly reviewing the science and vocabulary of artificial intelligence (AI), we offer an account of the intuition grounded in equality principles. Boiled down, our argument is that in a liberal democracy, there must be an aspect of 'role-reversibility' to certain judgments. In some contexts, those who exercise judgment should be vulnerable, in reverse, to its processes and effects. And those subject to its effects should be capable, reciprocally, of exercising judgment. In other words, there is, and ought to be, a sense in which the participants' roles in the process could always be inverted: in a different world, but for a contingent series of events, the decision-maker could be in the vulnerable position, not the powerful one. The fewer the contingencies in that inversion, the more role-reversible the decision.

Role-reversibility enables decision-makers to respect the gravity of decision-making from the perspective of affected parties. This, in turn, allows the act of judgment to be understood as the vindication of values shared by a broader moral community—a community of equals that includes both the decision-maker and the affected party, as well as many other people who were not involved in the decision but equally might have been, and who, in any case, share responsibility for the decision's consequences.

The central idea is that the act of applying a rule—the exact same rule—is different when carried out by a party who is herself subject to the rule as opposed to a party immune from the rule's reach. In a democracy, citizens do not stand outside the process of judgment, as if responding, in awe or trepidation, to the proclamations of an oracle. Rather, we are collectively responsible for each judgment. Thus, the party charged with exercising judgment—who could, after all, have been any of us—ought to be able to say: *This decision reflects constraints that we have decided to impose on ourselves, and in this case, it just so happens that another person, rather than I, must answer to them.* And the judged party—who could likewise have been any of us—ought to be able to say: *This decision-making process is one that we exercise ourselves, and in this case, it just so happens that another person, rather than I, is executing it.*

Questions of role-reversibility, then, are not spurred exclusively by powerful AI. To the contrary, democratic legitimacy has always been a problem of popular sovereignty: casting judgment as a joint enterprise over

which all members of the community can claim authorship.[4] For this to happen, it is not enough for rule-making processes to be inclusive, or for people to flock to voting booths on election day (though this would be wonderful to see). Acts of judgment, carried out in the name of the community, must bear the seal of collective application; although every act of judgment is particular, applying in *this* specific case to *this* specific party, all such acts should, at some level, apply to all of us. Thus, even if they apply the same criteria and reach the same outcomes, it is fundamentally different for a king or queen, standing above the law, to cast judgment on one of their subjects, or for the high-born, in a caste system, to decide the fate of the low-born.[5] And for the same basic reason, it would be fundamentally different for a machine to have the ultimate say over decisions. The rules and outcomes may be functionally identical, but the acts of deciding would be different. They would not be democratic acts.

After defending these principles in the abstract, we turn, in Part III, to a paradigm case: criminal judgment. We are not yet certain about—and we take no definitive position about—exactly *which* decisions depend on role-reversibility for their democratic legitimacy. Nor are we certain about the precise contexts in which, even if role-reversibility is not strictly necessary for legitimacy, dispensing with it should nevertheless require strong countervailing reasons. We strongly suspect, however, that decisions of criminal justice likely fall within one of these two camps, meaning—at the very least—that we ought to think long and hard before abandoning role-reversibility in the criminal context. According to the principle's dictates, then, those who sit in criminal judgment should be meaningfully vulnerable to the same criminal justice processes as criminal defendants. And, likewise, criminal defendants should be capable of exercising judgment. But for the recent turn of events, the parties' roles might just as well have been swapped: the criminal defendant could have been wearing the robes or sitting in the jury box, and the judge or jurors hauled to court as defendants.

This, we will argue, is the normative basis—if not necessarily the historical foundation—of the Anglo-American 'jury of peers' ideal, and it speaks to why the jury has long been celebrated as an organ of 'folk wisdom.' The idea is not that jurors have a better sense of right and wrong

---

[4] *See* Kiel Brennan-Marquez & Paul W. Kahn, *Statutes and Democratic Self-Authorship*, 56 WM. & MARY L. REV. 115, 173–77 (2014).

[5] Needless to say, the functional similarities between the political-economic conditions of the United States and a traditional caste system—and whether, by implication, certain of our institutional structures of judgment may already be offensive to the role-reversibility principle, quite independent of the machine question—is an issue prompted, not resolved, by our analysis.

than institutional actors do. (Though that may also be true.) It is, more fundamentally, that jurors respond to the act of judgment *as citizens*, not as officials, and in this respect, jury trials are a model of what role-reversibility makes possible: even when a jury trial does not lead to a different outcome than a trial before an institutional judge (or other fact-finding process), it facilitates the systemic recognition of judgment's human toll. And even more fundamentally, it transforms the trial into a democratic act.

One appealing feature of our account—even putting all other benefits to one side—is that it resists the 'humanity-fetishism' (or 'speciesism') that often looms over conversations about humans and machines. There is nothing special, in our view, about a human decision-maker. Rather, what matters is whether the decision-maker could swap positions with the affected party. In this sense, our account provides a ready-made answer for when it could become normatively acceptable for robots to don judicial robes, serve on juries, and occupy other democratic decision-making roles: when they interchangeably become robo-defendants.

•

Three caveats before jumping in. First, by describing role-reversibility as a principle whose value does not hinge on identifying and reducing errors, we do not mean to imply that role-reversibility *cannot* do this. It certainly might. Indeed, there may be reason to think that role-reversibility can inspire decision-makers to deviate from prescribed rules—as when juries nullify—in potentially salutary ways. We leave that question to future work. For the moment, our claim is simply that even assuming role-reversibility will not improve the accuracy of decision-making, it still has intrinsic value.[6]

Second, even if we are right about role-reversibility's intrinsic value, it does not follow that all other considerations, such as efficiency and accuracy, must go by the wayside. As with any complex normative question, there are likely to be competing dimensions of value when deciding whether to keep humans in the loop. For example, one can imagine a future in which the accuracy gains associated with machine decision-making are, at least in some settings, sufficiently great to justify dispensing with role-reversibility, and of course the tradeoff analysis may differ depending on what type of decisions—in terms of frequency,

[6] An analogy might be to the Constitution's guarantee of counsel in criminal cases, a guarantee that almost certainly helps avoid errors. But even if it did not—even if, for example, studies came to light suggesting that counsel makes little difference to trial outcomes—one still might argue that equal-access legal representation is an intrinsic good, a commitment to procedural integrity and human dignity regardless of outcomes.

complexity, and gravity—are involved. The upshot is simply that role-reversibility is a good as such, and likewise that its diminution (or elimination) would be a loss as such. In this sense, role-reversibility provides a conceptually sturdy foundation, in domains where the democratic pedigree of decision-making matters, for insisting on the value of human judgment—not insofar as human judgment leads to superior results (in fact, it may not), but insofar as humans are the ones being judged.

Third, by lionizing role-reversibility, we do not mean to suggest that all, or even most, decision-making in existing legal systems satisfies the normative criterion in practice. Far from it. Though we save it for future work, the question of whether the contemporary United States—in virtue of its class stratification and the racial disparities of its criminal justice system—fails role-reversibility 'across the board' is an important one, and one directly prompted (though obviously not resolved) by our analysis. Powerful AI is our motivating example. But when it comes to practical implications for the organization of our legal system, AI might be merely the tip of the iceberg—or, if you like, a canary in the mine.

I. AI AND HUMAN JUDGMENT

Given our thesis—which assumes technology will progress far beyond its current form—there is no need to belabor the current state of artificial intelligence. Nevertheless, a brief recitation of where the science has been, where it is going, and how long its goals might take, will dispel any concern that we are posing a fanciful question. Now is the time, before science in its natural course provides a coup de grâce, to begin considering what normative and legal rules should govern our brave new world.[7]

A. ARTIFICIAL GENERAL INTELLIGENCE

Artificial intelligence is a vast field of study, adherent to no single dogma.[8] And as with any discipline, definitional quandaries have persisted from its earliest days. For example, the pioneering work of Alan Turing provided the 'Turing test,' which asks whether a human interrogator, after

[7] For a version 1.0 of some of those rules see Stephen E. Henderson, *A Few Criminal Justice Big Data Rules*, 15 OHIO ST. J. CRIM. L. 527 (2018).

[8] In the words of the leading textbook, AI "encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; and everything from microelectronic devices to robotic planetary explorers." STUART J. RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH vii (3d ed. 2010).

interacting with an AI system, would mistake it for a fellow human.[9] This test captured the imagination of a generation of computer scientists, and even some philosopher-poets (as Brian Christian chronicles in his wonderful book, *The Most Human Human*).[10]

But the foundational puzzles of AI long predate Turing; in fact, they trace back at least as far as the origins of Western philosophy in ancient Greece. To the extent that Plato or Aristotle was correct about what constitutes rational thought, they arguably defined the goal of AI.[11] On the other hand, is that goal to build machines that (a) think or (b) act rationally? Or is the goal to build machines that (a) think or (b) act like humans? Experts disagree.[12] One thing, however, is clear: arriving at the "holy grail" of human-level and human-breadth intelligence, often termed "artificial general intelligence" or AGI, will require further developments in a wide range of disciplines.[13] Today's 'intelligent' machines are extremely good at particular tasks, like playing chess or flying planes. Humans, by contrast, are broadly good at many things, from programing those computers to composing music to (hopefully) landing planes when flight plans or systems go awry. At what point, then, might we expect to see genuine AGI?

Again, experts disagree. Colorable views range from the "techno-skeptics," who think we will not reach AGI for a century, if ever, to the

---

[9] *See id.* at 2. Think of Google's Duplex (impressively yet creepily) scheduling an appointment at a hair salon. *See* Gary Marcus & Ernest Davis, *A.I. Is Harder Than You Think*, N.Y. TIMES (May 18, 2018), https://www.nytimes.com/2018/05/18/opinion/artificial-intelligence-challenges.html [https://perma.cc/PA8Y-CLYR].

[10] BRIAN CHRISTIAN, THE MOST HUMAN: WHAT ARTIFICIAL INTELLIGENCE TEACHES US ABOUT BEING ALIVE (2011). A ready criticism of the Turing test is that *mimicry* is not *being*: "Aeronautical engineering texts do not define the goal of their field as making 'machines that fly so exactly like pigeons that they can fool even other pigeons.'" RUSSELL & NORVIG, *supra* note 8, at 3. Quite right.

[11] RUSSELL & NORVIG, *supra* note 8, at 4, 5.

[12] *See id.* at 2–5; *see also* TEGMARK, *supra* note 1, at 49–55.

[13] RUSSELL & NORVIG, *supra* note 8, at 27; TEGMARK, *supra* note 1, at 30, 39, 52 ("The holy grail of AI research is to build 'general AI' (better known as *artificial general intelligence*, AGI) that is maximally broad: able to accomplish virtually any goal, including learning."); *see also, id.* at Table 1.1. Nick Bostrom describes AGI like this: "Machines matching human intelligence—that is, possessing common sense and an effective ability to learn, reason, and plan to meet complex information-processing challenges across a wide range of natural and abstract domains." NICK BOSTROM, SUPERINTELLIGENCE 4 (2014). Russel and Norvig stress that this will require developments in mathematics (from computability to completeness to probability), economics (from utility to game theory to satisficing), neuroscience (how do our hundred billion neurons get the job done?), psychology (from behaviorism to cognitive science), and computer engineering, control theory, cybernetics, and linguistics. RUSSELL & NORVIG, *supra* note 8, at 5–16.

optimistic few who think it might happen very soon.[14] Most believe we are looking at somewhere between a few decades to a century.[15] Either way, the upshot is that if we want robo-jurors (or the equivalent), we may well have them within our lifetime or soon after.

•

We should note one important caveat: it is possible that the 'intelligence explosion' will achieve, but then almost immediately move past, AGI. Many people expect AGI to be ushered in by the 'singularity,' meaning the point at which humans can step aside because machines will take to improving machines.[16] (For *Hitchhiker* fans, think Deep Thought designing the Earth, only a whole lot faster.[17]) If, say, machines achieve AGI one day only to reach vastly superior intelligence the next, our machine overlords may have much more to say about our systems of criminal justice than we had planned. Yet any objection along these lines—whether an objection to our argument or to any other regarding the relationship between humans and intelligent machines—would prove too much. Is it possible that superior machines, once they arrive, will quickly take over just about *everything*? Certainly, and if that comes to pass, perhaps what has been previously theorized on a normative level will prove unenlightening, or at least irrelevant. But it may not come to pass; and if not, the proper relationship between human and machine capabilities will be crucial to resolve. So, onward we plow.

---

[14] *See* TEGMARK, *supra* note 1, at 30–33, 40–42. As Bostrom notes, the history of artificial intelligence is replete with overenthusiastic claims, and, often enough,

> [t]wo decades is a sweet spot for prognosticators of radical change: near enough to be attention-grabbing and relevant, yet far enough to make it possible to suppose that a string of breakthroughs, currently only vaguely imaginable, might by then have occurred. . . . Twenty years may also be close to the typical duration remaining of a forecaster's career, bounding the reputational risk of a bold prediction.

BOSTROM, *supra* note 13 at 4–5. For more on the historical 'ups and downs' of AI research, see RUSSELL & NORVIG, *supra* note 8, at 16–28 (providing a detailed timeline); TEGMARK, *supra* note 1, at 40–42 (describing "timeline myths"); BOSTROM, *supra* note 13, at 6–14 (describing "seasons of hope and despair").

[15] *See* TEGMARK, *supra* note 1, at 30–33, 40–42; BOSTROM, *supra* note 13, at 22–25.

[16] *See* TEGMARK, *supra* note 1, at 54, 134–35; BOSTROM, *supra* note 13, at 3–6. Two early waypoints of particular relevance to lawyers have been machine learning systems outperforming attorneys in certain tasks (especially in document review) and correctly predicting judicial outcomes. *See* Erin Winick, *Lawyer-Bots Are Shaking Up Jobs*, MIT TECH. REV., Dec. 12, 2017 (document review); Nikolas Aletras et al., *Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective*, 2 PEER J. COMPUTER SCI., Oct. 24, 2016.

[17] *See* DOUGLAS ADAMS, THE HITCHHIKER'S GUIDE TO THE GALAXY 181–83 (1st ed. 1980) (Chapter 28).

B. HUMANS IN THE LOOP

Whatever the precise timeframe of AGI, any machine of human-level intelligence will remain fallible for all the same reasons humans themselves are fallible; and any machine *approaching* human-level intelligence will likewise be fallible, *a fortiori*. Thus, the common view is that we ought to keep humans 'in the loop,' exercising ultimate say over the decision-making process.[18]

Arguments in favor of keeping humans in the loop take different forms, but they typically come back to the idea that humans understand decision-making *goals* and have broad intuitions, which enables them to identify problems or errors that elude machines.[19] To use a dramatic example, consider Stanislav Petrov's decision, on September 26, 1983, to override a Soviet system that mistakenly detected a nuclear attack.[20] Receiving notification from satellite-linked computers that a U.S. missile was incoming, Petrov's gut told him something was wrong, so he reported a fault in the system.[21] And when similar notifications kept appearing—for a second, third, fourth, and fifth nuclear missile, all of which were set to arrive within twelve minutes—Petrov's gut told him the same thing, so he once again reported a fault.[22] Petrov was right: the satellites had been confounded by solar reflections, though he was completely unaware of that detail at the time.[23] He simply *felt* something was 'off,' and that feeling may have avoided a disastrous nuclear war.[24]

---

[18] There is also the possibility that a machine intelligence might be *evil*, not sharing the normative values of humanity. *See, e.g.*, BOSTROM, *supra* note 13, at 130 (proposing an "orthogonality thesis" that claims "[i]ntelligence and final goals are orthogonal: more or less any level of intelligence could in principle be combined with more or less any final goal"). We are not so sure. In any event, this eventuality would so obviously dictate against replacing human judgment with any such machine that we do not belabor it.

[19] For a superb philosophical exposition of this view, see HUBERT DREYFUSS, WHAT COMPUTERS STILL CAN'T DO (1992).

[20] *See Midnight and Counting*, ECONOMIST, Sep. 30, 2017 at 86.

[21] *Id.*

[22] *Id.*

[23] *Id.*; *see also* TEGMARK, *supra* note 1, at 112–13.

[24] Another such Soviet hero is Vasili Arkhipov, who refused to launch a nuclear torpedo during the Cuban Missile Crisis. *See* Yasmeen Serhan, *When the World Lucked Out of a Nuclear War*, ATLANTIC (Oct. 31, 2017), https://www.theatlantic.com/international/archive/2017/10/when-the-world-lucked-out-of-nuclear-war/544198/ [https://perma.cc/JRY4-4RYM]; *see also* TEGMARK, *supra* note 1, at 112. Tegmark's *Future of Life Institute* posthumously honored Arkhipov in late 2017. *See* Tucker Davey, *55 Years After Preventing Nuclear Attack, Arkhipov Honored With Inaugural Future of Life Award*, FUTURE OF LIFE INSTITUTE (Oct. 27, 2017), https://futureoflife.org/2017/10/27/55-years-preventing-nuclear-attack-arkhipov-honored-inaugural-future-life-award/ [https://perma.cc/Q5G8-THKQ].

Cases like this—as well as their tragic counterparts, in which machines, absent human intervention or override, have made disastrously wrong choices[25]—have led many commentators in AI policy circles to argue that humans should *always* be kept 'in the loop,' especially when it comes to grave and irreversible decisions like the use of weapons systems.[26] Similar arguments have emerged in the context of law enforcement,[27] medicine,[28] finance,[29] and the administration of social programs.[30] In

[25] For example, five years after the episode with Petrov, a similarly erroneous notification came in to the USS Vincennes guided missile cruiser, but no one second-guessed the machine—leading to a 'retaliatory' strike that destroyed Iran Air Flight 655, killing all 290 persons on board and galvanizing decades of international distrust. *See* Max Fisher, *The Forgotten Story of Iran Air Flight 655*, WASH. POST, Oct. 16, 2013, at A13; *see also* TEGMARK, *supra* note 1, at 111.

[26] *See* TEGMARK, *supra* note 1, at 111–13; Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO L. REV. 1837, 1872–83 (2015) (summarizing a variety of arguments to the same effect). For more general arguments along the same lines, see, for example, Bryan Casey et al., *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise* (forthcoming) (explaining the European Union's 2016 "right not to be subject to a decision based solely on automated processing," including a right "to obtain human intervention"); Tamara Capeta, *Of Judges and Robots* in CHALLENGES OF LAW IN LIFE REALITY 129–42 (2017) (arguing that robots lack necessary "emotion"); A. Michael Froomkin et al., *When AIs Outperform Doctors: The Dangers of a Tort-Induced Over-Reliance on Machine Learning and What (Not) To Do About It* (forthcoming) (arguing that machine learning health algorithms without a human in the loop would lead to a decrease in the quality of care).

[27] With regard to policing, see, for example, Michael L. Rich, *Machine Learning, Automated Suspicion Algorithms, and the Fourth Amendment*, 164 U. PA. L. REV. 871, 897–98 (2015) (arguing against the automation of reasonable suspicion on the grounds that "human beings are always at least potentially capable of including a new piece of relevant information in the analysis," while "database[s] cannot contain all the facts that are relevant in every case," which means that machine learning algorithms "cannot consider the 'whole picture' regarding a person's potential criminality as required by the Fourth Amendment"). With regard to sentencing, see, for example, Sonja Starr, *Evidence-Based Sentencing and the Scientific Rationalization of Discrimination*, 66 STAN. L. REV. 803 (2014). *See generally* Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Standards in the Age of Powerful Machines*, 70 VAND. L. REV. 1249 (2017) (arguing that explanatory standards in the criminal justice system defy automation, because humans have to remain 'in the loop' to explain *why* a decision is rendered as it is).

[28] *See, e.g.*, Andreas Holzinger, *Interactive Machine Learning for Health Informatics: When Do We Need A Human in the Loop*, 3 BRAIN INFORMATICS 119 (2016). A striking example of the value of human judgment in medicine is optimal triage for pneumonia patients. Back in the late 1990's, an algorithm designed to sort pneumonia patients into higher and lower risk categories was sending patients with asthma home rather than recommending hospitalization. The reason for this counterintuitive (and wrong) suggestion is because, historically speaking, pneumonia patients with asthma, being at greater risk, tended to receive more care. Thus, they had better overall health outcomes. These outcomes were 'all the algorithm saw,' so to speak, and so it sorted the asthmatic pneumonia patients

addition to these 'disaster-focused' arguments, some observers have emphasized other benefits associated with human oversight—for example, that affected parties may have an easier time respecting hard-to-swallow outcomes when they know that a human took part in the decision-making process;[31] and that humans may be able to identify patterns that, once parsed, can be 'reinvested' into the decision-making system and refine outcomes over time.[32]

We do not minimize any of these instrumental arguments in favor of human judgment. They are certainly valid today, and they may survive the next generation of AI. The rest of this article is dedicated, however, to exploring what should happen if arguments like these do *not* survive. If technology evolves in ways that effectively dislodge the human claim to superiority in appreciating the goals of decision-making—and in identifying certain classes of errors—can human judgment still be defended?

---

as lower risk. Only through human oversight was this error detected and solved. *See* Paul Voosen, *How AI Detectives are Cracking Open the Black Box of Deep Learning*, SCIENCE (July 6, 2017).

[29] *See, e.g.*, Yesha Yadav, *How Algorithmic Trading Undermines Efficiency in Capital Markets*, 68 VAND. L. REV. 1607 (2015). A cautionary tale is the 2010 "Flash Crash," in which automated stock trading algorithms caused a trillion-dollar stock-market crash by acting in a manner incomprehensible to a human trader. *See* BOSTROM, *supra* note 13, at 20–21. In a twist, however, it was another automated system that actually caught the problem. *See id.*

[30] *See, e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1256–57 (2008) (enumerating algorithmic governance tools that have been prone to error, including (1) "[b]enefit [m]anagement [s]ystems" that have issued "hundreds of thousands of incorrect Medicaid, food stamp, and welfare eligibility determinations," (2) algorithms meant to locate " 'dead-beat' parents who owe child support" that sweep in many non-offenders, triggering automatic garnishment of wages, and (3) counterterrorism tools that, due to "[u]nsophisticated algorithms and faulty data," end up "generat[ing] high rates of false positives" with grave law enforcement consequences).

[31] *See, e.g.*, TEGMARK, *supra* note 1, at 106–07 (suggesting that without explainable AI, affected parties may not "feel that they understand its logical reasoning enough to respect its judgment"); Ric Simmons, *Big Data and Procedural Justice: Legitimizing Algorithms in the Criminal Justice System*, 5 OHIO ST. J. CRIM. L. 573 (2018) (arguing that too much machine decision-making—at least in the criminal context—may adversely impact people's perception of "procedural justice").

[32] *See, e.g.*, Brennan-Marquez, *supra* note 277, at 1280–97. A common cause of this dynamic is 'overfitting,' which often requires human interpretation to identify (and, *a fortiori*, to rectify). In their book *Algorithms to Live By*, Brian Christian and Tom Griffiths include a memorable case in which a nine-factor model fits life-satisfaction data almost perfectly, but it absurdly predicts a dramatic rise from almost zero happiness just before marriage, and a similar plummet to zero happiness eleven years thereafter. *See* BRIAN CHRISTIAN & TOM GRIFFITHS, ALGORITHMS TO LIVE BY: THE COMPUTER SCIENCE OF HUMAN DECISIONS 152–53 (2016).

II. DEMOCRATIC EQUALITY AS ROLE-REVERSIBILITY

Imagine, as we did at the outset, a world in which machine-only decisions are more accurate and consistent, across cases, than human-plus-machine decisions.[33] In that sort of world, would there be any value in keeping humans in the loop? Would there be grounds to insist on human judgment as an essential aspect of certain kinds of decision-making, quite independent of—and even in *spite* of—the outcomes it begets?

We think so. And in this section we explore why, focusing on the concept of 'role-reversibility' in decision-making—the idea that those who exercise judgment should be reciprocally vulnerable to its processes and effects. In other words, there is, and ought to be, a sense in which participants' roles in the process could always be inverted: in a slightly different world, but for a contingent series of events, the decision-maker could be the one judged, rather than the one doing the judging.

What matters, then, is not the fact of humanness per se. What matters is whether decision-makers are situated to imagine themselves into the role of an affected party, and vice versa—such that both participants, and in some sense the entire moral community, can understand judgment as a democratic act.

A. 'THERE BUT FOR THE GRACE OF GOD'

The idea that one should be humbled by the possibility that life could have gone very differently—and in particular, that one could easily have occupied the position of 'judged'—strikes us as ancient. It may be as old as the human race. The modern version of the idea traces to the sixteenth century, when the English reformer John Bradford allegedly declared, after seeing convicts hung in the gallows, "but for the grace of God there goes John Bradford."[34] Much more recently, in 1964, folk musician Phil Ochs articulated the same sentiment without the religious overtone: "There but for fortune go you or I."[35]

---

[33] As we said at the outset, we realize there are many reasons to be skeptical of these premises coming to fruition any time soon—or maybe ever. But we are consciously pitching this argument *in principle*, on the assumption that the instrumental values of human judgment have disappeared. To whatever extent those values are still operative in practice, they "put[ ] extra icing on a cake already frosted." Yates v. United States, 135 S. Ct. 1074, 1093 (2015) (Kagan, J., dissenting).

[34] THE WRITINGS OF JOHN BRADFORD, M.A. xliii (Aubrey Townsend ed., 1853), https://archive.org/stream/writingsofjohnbr02braduoft#page/n5/mode/2up [https://perma.cc/HA8J-TKM6].

[35] PHIL OCHS, *There But For Fortune*, *on* NEW FOLKS VOL. 2 (Vanguard 1964).

The sentiment reverberates. The claim is that our similarities dwarf our differences; that life depends largely on fortune or, in religious parlance, grace. If the winds of fortune blew different, many of our individual lives—and at some level, all of social life—would be different as well.

What makes this idea powerful, and powerfully tied to democracy, is that it binds together our fates. It is the theological precursor to John Rawls' theory of justice, T.M. Scanlon's theory of fairness, and other philosophical systems that insist on resolving normative questions *out of context*—without regard to which particular people stand to lose or gain.[36] Rawls famously invoked the "veil of ignorance" to capture this idea,[37] and Scanlon developed an analogous account of "generic reasoning."[38] Either way, the upshot is the same: distributive decisions are only legitimate if they can be justified across the board, irrespective of who happens to occupy the 'winning' and 'losing' positions.

As applied to the formulation of rules—resolving systems-level questions in ways that advance the general interest, not the interests of particular groups—this principle is quite familiar. Indeed, it verges on self-evident. In a democratic society, we naturally gravitate toward 'generic' justifications when debating the merits of different policies.[39] Or, put the other way around, justifications that do *not* take a generic form feel intuitively out of keeping with democratic sensibilities. If someone tries to

---

[36] T.M. SCANLON, WHAT WE OWE TO EACH ANOTHER 189–90 (1998) (describing different philosophical positions—including his own, as well as Rawls's, Kant's, and Habermas's—that embrace some version of 'out of context' reasoning about moral questions).

[37] *See* JOHN RAWLS, A THEORY OF JUSTICE (1971); *see also* John E. Roemer & Alan Trannoy, "*Equality of Opportunity*," *in* HANDBOOK OF INCOME DISTRIBUTION (Atkinson & Bourgoignon eds., 2014) (explaining that Rawls's central ambition was to argue for "shield[ing] decision-makers from knowledge of information about their situations that [is] 'morally arbitrary,' so that the decisions they [come] to regarding just allocation would be impartial . . . [thus] deriv[ing] principles of justice from rationality and impartiality").

[38] *See* SCANLON, *supra* note 36; Rahul Kumar, *Reasonable Reasons in Contractualist Moral Argument*, 114 ETHICS 6, 6–7 (2003) (summarizing Scanlon's point as follows: the justifiability of regulation turns on "whether or not a particular proposed principle for the [] regulation . . . is one that no one, suitably motivated, could reasonably reject as a basis for informed, unforced, general agreement").

[39] In practice, of course, the benefits of our design choices will accrue to some groups rather than others; in reality, all of us *do* occupy specific social positions, even if we try to generalize beyond those positions when making normative arguments. Indeed, the task of persuasion is very often to convince someone else to abstract away from their own social position—to embrace a normative viewpoint contrary to their own immediate interests (or, likewise, to justify a normative viewpoint that is favorable to one's immediate interests by establishing that the *reason* one embraces the viewpoint is generic, not personal).

defend a policy on the grounds that it will line her own pockets, or that it will disadvantage a particular group, the argument will not just be unconvincing—it should not even merit consideration.

Yet the process of reasoning 'generically,' of taking a 'there but for fortune' approach to judgment, also bears on the application of rules to particular cases—if anything, even more acutely. When it comes to the formulation of rules, all members of the community are 'reversible,' such that everyone ought to be able to say: *the rationale behind this rule is one that I could be expected to embrace, regardless of my social position*. Of course, this is a very broad concept, requiring an imaginative feat encompassing infinite distributive worlds. Yet this style of reasoning is also quite natural. Precisely *because* of its breadth and how difficult it can be to predict overall welfare effects—which specific people will end up, when all is said and done, bearing the burdens of a given policy—we often gravitate toward generic reasoning by default.

By contrast, in the application of rules to particular cases, it is quite clear which specific people bear the relevant burdens—the party called on to decide, who must endure the "agony of judgment,"[40] and the party whose fate the decision will ultimately shape. Moving from the abstract to the particular, 'role-reversibility' can become more than simply a metaphorical vehicle of reasoning. It can become a psychological reality, allowing participants in the decision-making process to *literally* imagine switching places. And when that happens, the participants are forced, at some level, to occupy the 'generic' position.

Put a bit more poetically, one can imagine the affected party—in the event of an adverse decision—thinking something along the following lines:

> Obviously, I wish I could avoid this. I don't want this stigma, this pain. Still . . . I cannot say, that had I been hauled into the role of juror—and, really, but for a few contingent circumstances I could have been—I would have done any differently or 'better.' In that sense, I suppose, I have *some* measure of acceptance. At the very least, this decision-maker—my particular 'judge'—isn't to blame.

Likewise, in reverse: one can imagine the decision-maker—the person called on to judge a member of her own moral community[41]—engaging in a thought-process like so:

---

[40] Owen M. Fiss, *Against Settlement*, 93 YALE L. J. 1073, 1086 (1984).

[41] *Cf.* Matthew 7:1–2 ("Judge not, that ye be not judged. For with what judgment ye judge, ye shall be judged: and with what measure ye mete, it shall be measured to you again."); Avot 2:4 ("And do not judge your fellow until you have stood in his place.").

> I didn't ask for this; really, I'd rather be doing most anything else. I don't claim any moral—or other—superiority. But for a few contingent circumstances, it could have been my fate, rather than his, on the line. And if it were, if the tables were turned, he'd presumably make the same judgment about me, and I'd have to accept the same outcome. So, I suppose I can walk away from this. Not because I did 'right,' whatever that means. But because I believe he—the judged party—would have done the same.

In this sense, the act of applying rules is a more direct reminder of social contingency than the (more rarefied) act of formulating rules. When 'role-reversible' decision-makers exercise judgment in particular cases, they come face to face with the reality that, but for a largely random series of events, they might have been in the affected party's shoes. In a meaningful sense, they are judging themselves.

B. JUDGMENT AND SELF-AUTHORSHIP

Such generic reasoning—both in the formulation of rules and in their application to particular cases—is integral to democracy. To be democratically legitimate, rules and judgments should be the outcome of popular sovereignty, something over which the people exercise authorship. Voting is an element of this, but not sufficient by itself.[42] Long after we get home from the ballot box, the question is whether we can accept both the formulation and application of rules—statutes, policies, and other official acts—as our own.[43]

At a systemic level, the question is whether rules can be understood as *constraints we impose on ourselves* rather than constraints imposed on us from without. It is this ideal that distinguishes democracies from monarchies, caste systems, and colonial arrangements. To satisfy the ideal, not everyone has to agree on the content of all rules at all times—a completely impracticable ambition. But democratic pedigree still matters. Rules do have to trace, ultimately, to generic reasons; they have to be about more than rent-seeking, nepotism, plutocracy, and other naked assertions of power.

And the same is true, for the same fundamentally democratic reason, of individual judgments. Just as we should all be able to see rules as constraints we impose on ourselves, decision-makers should be able to see

---

[42] Hence the appeal of 'fiduciary political theory'—the notion that lawmakers, judges, and other public officials are best understood as operating as agents of the people. *See, e.g.*, Ethan J. Lieb et al., *A Fiduciary Theory of Judging*, 101 CAL. L. REV. 699 (2013); D. Theodore Rave, *Politicians as Fiduciaries*, 126 HARV. L. REV. 671 (2013).

[43] *See* Brennan-Marquez & Kahn, *supra* note 4, at 115, 163–77. In political theory, this idea ultimately traces back to Hobbes. *See id.* at 163–73; THOMAS HOBBES, THE LEVIATHAN (1668).

the act of judgment as something they are doing, in effect, to themselves. Not literally, for of course it is the affected party, not the decision-maker, who must bear the burden of judgment. But figuratively and aspirationally—since the decision-maker could have just as easily (but for fortune) been in the position of the affected party, and the basis for the particular judgment, just like the basis for the rule from which it flows, ought to depend on reasons that have nothing to do with social contingency.

In this sense, the stylized forms of reasoning imagined above—whereby an affected party contemplates rendering a decision, and a decision-maker contemplates bearing its burden, reciprocally—could be elaborated even further. It is not merely that a decision-maker should be able to imagine accepting the same judgment in reverse. It is that she should be able to say: *this decision is an outcome of democracy; it reflects a constraint we have decided to impose on ourselves, and in this case, it just so happens that another person, rather than I, must answer to it*. Likewise, the affected party should not merely be able to imagine rendering judgment in reverse, but be able to say: *this decision is an outcome of democracy; it reflects a constraint we have decided to impose on ourselves, and in this case, it just so happens that another person, rather than I, is the one executing it.*

In other words, individual acts of judgment, no less than generalized rules, should be understandable as self-imposed. Although it happens to be *this* particular decision-maker who must render judgment, and *this* affected party who must endure its outcome, it could have been—and in a counterfactual world, may well have been—any of us, in either role. This, once again, is what separates democratic government from its more rigidly hierarchical alternatives. In practice, we know that formally democratic institutions often fall short of this principle, in which case decision-makers start to look much like royals or imperialists. But practical shortcomings do not drain the principle of significance. It remains an aspiration—a normative lodestar—all the same.

Of course, we also hope that democratic structures will produce welcome outcomes. Part of the argument in favor of democracy—historically, and often conceptually—is that it subjects power to popular accountability and, in doing so, makes power less susceptible to abuse. Whether this is true as an empirical matter may depend on whom one asks. But the important point, for our purposes, is that democracy's value does not actually depend on the outcomes it produces. Horrific outcomes may, as with anything, inspire a loss of faith. But the fundamental claim is not that democracies always perform better—always make superior decisions about rules—than alternative systems of government. Rather, it is a claim

about self-rule. Even when democratic outcomes fall short at a policy level, they are *ours*.[44]

The same goes for particular judgments. As with democratic lawmaking writ large, we may hope that role-reversibility would facilitate better decisions—if not in terms of concrete outcomes (since we are holding that variable constant for the sake of analysis),[45] at least in terms of how judgment is rendered: its tone, its texture, its regard for judgment's sometimes-tragic cast. It seems plausible, for example, that a role-reversible decision-maker would be likely to *agonize* over hard choices in a way that a decision-maker immune from the rule's reach would not. It also seems plausible that a role-reversible decision-maker would be inclined to issue an apology or other words of humility and understanding alongside the formal judgment, or, depending on the type of case, perhaps to include words of condemnation alongside the formal judgment.[46]

In short, whether or not these peripheral aspects of role-reversible judgment matter (or even occur), they are not what justify role-reversibility. The bedrock justification for preferring role-reversible judgments is their democratic fidelity.

•

Before moving on, we wish to anticipate an objection—and hopefully, in responding, to further shore up our claim. The objection is this. *Why*, one might ask, *should it be worrisome, on democratic grounds, to delegate the implementation of rules to machines? It is one thing to argue that, as a human community, we should be able to choose our own fate. But once that fate is chosen—once the rules are prescribed—what is so infirm about*

---

[44] A similar claim might be made for the principle of autonomy. Perhaps it leads to better outcomes, but perhaps it does not. Either way, the decisions are *yours*.

[45] As we noted previously, this is not to say that role-reversibility necessarily *does not* affect decisional outcomes. In fact, it well might. In a companion paper, we plan to explore the proposition that role-reversible decision-makers are more likely than their non-role-reversible counterparts to deviate from prescribed rules and orthodoxy, with jury nullification as the paradigm case. In other words, it seems to us descriptively plausible—and plausibly a source of normative value—that a decision-maker who can appreciate the 'there but for the grace of god/fortune' aspect of judgment will be more inclined to regard the application of rules to hard cases as morally egregious. For now, though, the point is that nothing in our argument depends on role-reversibility exerting a pull on decisional outcomes. Whether it does or not, the claim about democratic legitimacy stands.

[46] For example, Judge Chin of the Southern District of New York gave a pretty remarkable speech when he sentenced Bernard Madoff, describing Madoff's crimes as acts of "extraordinary evil" that cried out for maximum punishment. *See* Benjamin Weiser, *Madoff Judge Recalls Rationale For Imposing 150-Year Sentence*, N.Y. TIMES, June 29, 2011, at A1. Whether this kind of community-minded punitive sentiment changes the act of judgment—and for better or for worse—is a question we leave to future work.

entrusting the execution to powerful AI? After all, such delegation has at least one clear benefit: it relieves the agony of decision-making.

This objection may feel familiar. To begin with, it tracks the distinction between legal and factual questions. The common wisdom is that the former require normative reasoning, while the latter only require empirical reasoning—prefiguring a plausible division-of-labor between humans and machines.[47] Furthermore, it is an argument that technology-enthusiasts have been known to make. In algorithmic governance circles, pro-machine arguments often center on the distinction between values and goals (over which humans should have final say) and implementation outcomes (which, the logic goes, could be delegated to machines).[48]

The problem with this objection is that its logic proves too much. If we are willing to take a non-democratic approach to the implementation of rules, why not take a non-democratic approach to their formulation? If we are comfortable—as the objection presumes—bestowing machines with authority to implement our values, why not also allow them to *craft* those values?

The answer cannot be technological. Even if current AI is incapable of 'crafting values,' we cannot count on that being true forever.[49] Rather, the answer must be about the importance of democratic pedigree. It just seems illegitimate—in the manner of a caste system—to have values crafted, and rules imposed, from without. For the reasons explored above, however, the application of rules to particular cases *also* implicates democratic legitimacy; and also, by extension, requires adherence to principles of self-authorship. Ultimately, then, delegating implementation power to intelligent machines makes no more (or less) sense than delegating such power to an exclusive subset of humans—like the priests

---

[47] *See, e.g.*, Emad H. Atiq, *Legal vs. Factual Normative Questions & The True Scope of* Ring, 32 NOTRE DAME J.L. ETHICS & PUB. POL'Y 101, 103 (2018) ("The dominant view amongst legal theorists is that the law/fact distinction tracks or maps on to the distinction between normative and empirical questions.").

[48] *See, e.g.*, Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 GEO. L.J. 1147 (2017) (developing an argument along these lines using the rulemaking/adjudication dichotomy from administrative law).

[49] *See infra* Part I. For the sake of argument, we are assuming machines at or past the point of artificial general intelligence (AGI), meaning they are capable of the many and varied tasks of the human brain. There is no consensus on what form this technology might take, because there is no consensus on what intelligence *is*. But any assertion that an intelligent machine will be incapable of 'x' because x 'feels human' leaves much to be desired. For example, imagine neuroscientists are able to unlock the mysteries of our neural-network brains, and computer scientists are able to precisely replicate them in silicon. Then such machines *will* be able 'do values,' at least as well as humans can.

and knights of yesteryear. *Could* we do this? Yes, of course we could. But democratic values counsel otherwise.

III. JURY TRIALS AS THE PARADIGM CASE

Having considered and defended role-reversibility in the abstract, let us consider a particular form of paradigmatically role-reversible judgment: the jury. In his famous Hamlyn Lecture on *Trial By Jury*, Sir Patrick Devlin leads with this:

> Trial by jury is not a subject on which it is possible to say anything very novel or very profound . . . . The English jury is not what it is because some lawgiver so decreed but because that is the way it has grown up. Indeed, its invention by a lawgiver is inconceivable. We are used to it and know that it works; if we were not, we should say that it embodies a ridiculous and impracticable idea.[50]

We disagree—sort of.

Devlin's is one of the most important works on the jury; we admire his contributions and find many of them agreeable. His historical analysis, for example, strikes us as impeccable: Devlin deftly winds from the first administrative juries (inquests) to presentment juries (the predecessor of our grand jury) to adjudicative juries.[51] No singular—let alone rational—mind was guiding those choppy tides of history, and it is ever-refreshing to read Devlin's critique of those who would bring mystical precision to the tortured path of historic reality.[52] Our argument is not a historical one.

---

[50] DEVLIN, *supra* note 2, at 3–4.

[51] *Id.* at 5–14.

[52] For example, consider the traditional number of jurors, twelve:

> Many romantic explanations have been offered of the number twelve—the twelve tribes of Israel, the twelve patriarchs, and the twelve officers of Solomon recorded in the Book of Kings, and the twelve Apostles. Not all of these suggestions are equally happy; the first implies that there may be a thirteenth juror who has got lost somewhere in the corridor and the last that there is a Judas on every jury. It is clear that what was wanted was a number that was large enough to create a formidable body of opinion in favour of the side that won; and doubtless the reason for having twelve instead of ten, eleven or thirteen was much the same as gives twelve pennies to the shilling and which exhibits an early English abhorrence of the decimal system.

Id. at 8. Or consider the longstanding rules that allowed judges to keep juries confined without food until they reached a unanimous verdict:

> All this is very reminiscent of the verdict's origin. What was sought was not a rational conclusion but a sign, something akin to the result of the ordeal or to triumph in battle; the process could not be determined until it was obtained and, once obtained, the methods of obtaining it were thought less important than the fact that it was there.

Id. at 51. For more on the move from ordeal to the jury trial, see Elizabeth Papp Kamali & Thomas A. Green, *The Assumptions Underlying England's Adoption of Trial by Jury for*

And we agree with Devlin's claim that some aspects of our juries remain rather "ridiculous and impracticable," or at least imperfect.[53] Our argument is not a paean to jury trials, especially as they tend to be practically implemented.[54] We take issue only with Devlin's claim that the jury embodies a "ridiculous and impracticable *idea*."

But even here, we are not really disagreeing with Devlin. There *is* much that seems impracticable, and even a little ridiculous, in a jury system.[55] So, perhaps the most accurate statement of our thesis is that we ought to preserve the jury—and particularly the criminal jury—*despite* its shortcomings. Why? Because juries robustly embody the democratic spirit

---

Crime in LAW AND SOCIETY IN LATER MEDIEVAL ENGLAND AND IRELAND: ESSAYS IN HONOUR OF PAUL BRAND (Travis Baker ed., 2018).

[53] For example, why it is that our juries offer no explanation nor justification for their verdicts, when we generally know that requiring a justification leads to a better-reasoned decision? In Devlin's mind, merely on account of historical anachronism: "It is the oracle deprived of the right of being ambiguous. The jury was in its origin as oracular as the ordeal—neither was conceived in reason—the verdict, no more than the result of the ordeal, was open to rational criticism." DEVLIN, *supra* note 2, at 14 (slight change in punctuation). Especially sympathetic cases are finally pushing on this centuries-long practice. *See* Pena-Rodriguez v. Colorado, 137 S. Ct. 855 (2017) (constitutionally requiring a state to permit impeachment of a jury verdict when there are accusations of racial prejudice during deliberations); *see also, e.g.*, Andrew E. Taslitz & Stephen E. Henderson, *Reforming the Grand Jury to Protect Privacy in Third Party Records*, 64 AM. U. L. REV. 195 (2014) (criticizing the lack of grand jury protections).

[54] For a thoughtful modern defense of our criminal jury and a citizen 'call to action' see ANDREW GUTHRIE FERGUSON, WHY JURY DUTY MATTERS: A CITIZEN'S GUIDE TO CONSTITUTIONAL ACTION (2012). Ultimately, Ferguson's arguments, like ours, defend the system of jury adjudication, not any failures in its implementation.

[55] Consider this wonderful description by Devlin:

> Twelve (why twelve?) men and women are to be selected at random; they have never before had any experience of weighing evidence and perhaps not of applying their minds judicially to any problem; they are often, as the Common Law Commissioners of 1853 tactfully put it, "unaccustomed to severe intellectual exercise or to protracted thought." The case may be an intricate one, lasting some weeks and counsel may have in front of them piles of documents, of which the jury are given a few to look at. They may listen to days of oral evidence without taking notes—at least, no one expects them to take notes and no facility is provided for it in the jury-box, not even elbow room. Yet they are said to be the sole judges of all the facts. At the end of the case they are expected within an hour or two to arrive at the same conclusion. Without their unanimous verdict no man can be punished for any of the greater offences. Theoretically it ought not to be possible to successfully enforce the criminal law by such means.

DEVLIN, *supra* note 2, at 4–5.

of role-reversibility.[56] In the words of one participant, the jury is "the purest example of democracy in action that I have ever experienced."[57]

A. A 'JURY OF YOUR PEERS'

In so defending the jury, we must separate the ideal from the historic. Juries have always been, to say the least, imperfect. They have delivered some of our greatest injustices,[58] and have been anything but role-reversible, excluding those without ample property or those of a disfavored race, sex, or religion.[59] Nonetheless, with time, the institution has become increasingly democratic, hewing closer to a role-reversible design.

The right to a criminal jury trial has impressive pedigree: it is the only provision of the American Bill of Rights that was expressly redundant, as Article III already provided that "[t]he trial of all Crimes . . . shall be by Jury."[60] As interpreted today, those rights require jurors drawn from a fair

---

[56] We think it possible that Devlin himself would have been sympathetic to this claim. He certainly believed that while the "origin [of the jury] is accidental . . . its retention [is] deliberate." *Id.* at 154. And while at times he urges a rather noncommittal, consequentialist argument for preserving the jury (*see, e.g.*, *id.* at 57), when one considers his entire argument, he might have been an early proponent of role-reversibility, albeit primarily for its deviation through nullification component. *See id.* at 155–65. We leave this fascinating aspect of jury decision-making to future work.

[57] Raymond J. Brassard, *Juries Help Keep Our Democracy Working*, BOSTON GLOBE, May 1, 2003, at A19 (quoting a juror letter written following a trial). The juror continued: "Everybody participated. I don't know if we arrived at the truth, but I felt we did a good job." *Id.* For a more theoretical account of the same point, see Jenny Carroll, *The Jury as Democracy*, 66 ALA. L. REV. 825 (2015).

[58] *See, e.g.*, Hector (A Slave) v. State, 2 Mo. 166, 166 (1829) (considering a jury conviction where the defendant confessed as a result of being "flogg[ed] all night, . . . scream[ing] under the lash"); Brown v. Mississippi, 297 U.S. 278, 281 (1936) (considering a jury conviction where the defendant confessed as a result of being "hanged . . . by a rope to the limb of a tree [twice], and [then] . . . tied to a tree and whipped"). Remarkably, in *Brown*,

> This deputy was put on the stand by the state in rebuttal, and admitted the whippings. It is interesting to note that in his testimony with reference to the whipping of [one] defendant . . . , and in response to the inquiry as to how severely he was whipped, the deputy stated, "Not too much for a negro; not as much as I would have done if it were left to me." Two others who had participated in these whippings were introduced and admitted it—not a single witness was introduced who denied it.

Id. at 284–85.

[59] *See* Albert W. Alschuler & Andrew G. Deiss, *A Brief History of the Criminal Jury in the United States*, 61 U. CHI. L. REV. 867, 876–901 (1994) (chronicling historic changes in the composition of the American jury).

[60] U.S. CONST. art. III § 2; *see also* U.S. CONST. amend. VI. As Alschuler and Deiss point out, this redundancy is no surprise—it is also the only right shared by all twelve state constitutions written prior to the federal Constitutional Convention. *See* Alschuler & Deiss, *supra* note 59 at 870. In the words of one district judge, the jury "is as American as rock 'n'

cross section of the community, no "distinctive group" systematically excluded.[61] Even the 'unrestricted' peremptory challenge no longer permits racial bias.[62] And jurors are reciprocally vulnerable to judgment's processes and effects: not only must they live in the same jurisdiction, subject to the same laws,[63] they can be criminally punished for inappropriate conduct in deciding the case before them.[64] Reciprocally, those subject to judgment—the criminal defendants—might have served on a past jury, and they remain capable of serving on a future one.[65]

At least *in theory*, then, the participants' roles in the jury trial process could readily be inverted. The counterfactual world in which any particular juror is, instead, sitting in the defendant's chair is hardly remote—but for a minor series of contingent events, it easily could have been the reality. This role-reversible dynamic is, we think, the conceptual foundation—albeit not the historical pedigree—of the 'jury of your peers' ideal. It explains why

---

roll." William G. Young, *Vanishing Trials, Vanishing Juries, Vanishing Constitution*, 40 SUFFOLK U.L. REV. 67 (2006).

[61] *See* Taylor v. Louisiana, 419 U.S. 522, 530 (1975) (holding that trial venires must constitute a fair cross section of the community, meaning no distinctive group is systematically excluded).

[62] *See* Batson v. Kentucky, 476 U.S. 79 (1986) (prohibiting racially based peremptory challenges).

[63] *See* U.S. CONST. amend. VI ("In all criminal prosecutions, the accused shall enjoy the right to a . . . trial, by an impartial jury of the state and district wherein the crime shall have been committed, which district shall have been previously ascertained by law.").

[64] *See, e.g.*, Molly McDonough, *Rogue Jurors*, ABA JOURNAL, Oct. 2006 (explaining, among other cases, an instance in which a juror was sentenced to six and a half years in prison for conspiracy, contempt, and obstruction of justice).

[65] The qualifications for federal jury service are representative and minimal:

> In making such determination the . . . judge . . . shall deem any person qualified to serve on grand and petit juries in the district court unless he—
>
> (1) is not a citizen of the United States eighteen years old who has resided for a period of one year within the judicial district;
>
> (2) is unable to read, write, and understand the English language with a degree of proficiency sufficient to fill out satisfactorily the juror qualification form;
>
> (3) is unable to speak the English language;
>
> (4) is incapable, by reason of mental or physical infirmity, to render satisfactory jury service; or
>
> (5) has a charge pending against him for the commission of, or has been convicted in a State or Federal court of record of, a crime punishable by imprisonment for more than one year and his civil rights have not been restored.

28 U.S.C. § 1865(b)(2015). Some additional persons will be exempt (28 U.S.C. § 1863(b)(6)(2015)) or excused (28 U.S.C. § 1863(b)(5)).

not even the criminal defendant can unilaterally waive jury review.[66] And while any single jury of twelve is hardly likely to be well (or at least fully) representative of one's 'peers,' the criminal jury embodies a democratic, role-reversible norm.[67] In the words of Alexis de Tocqueville, the American jury is, and was meant to be, "a political institution," "one form of the sovereignty of the people."[68]

B. THE VANISHING CRIMINAL JURY

It is little surprise, then, that concern about the disappearance of criminal jury trials is nothing new, at least if its replacement is likely—as practice suggests—to be a less role-reversible, less democratic system of plea bargaining seemingly governed more by concerns of efficiency (and even career-advancement) than by concern for properly deciding each individual case.[69] Over sixty years ago, the jury's decline was an impetus

---

[66] *See* Singer v. United States, 380 U.S. 24 (1965) (holding "that the Constitution neither confers nor recognizes a right of criminal defendants to have their cases tried before a judge alone"). According to John Langbein,

> We have historical records from the English sources of a few Eighteenth-Century cases in which some pathetic accused, caught in the act or otherwise sensing the hopelessness of his case, attempted to plead guilty. . . . Time and again the judge urged the accused to plead 'not guilty' and to take his case to the jury.

John H. Langbein, *On the Myth of Written Constitutions: The Disappearance of Criminal Jury Trial*, 15 HARV. J. L. & PUB. POL'Y 119, 120 (1992).

[67] While we believe this role-reversibility to be the linchpin of jury legitimacy, we note lay jurors also have an 'outside-the-system-ness' that can provide further benefits. In the words of Devlin, "[t]he jury hear the witness as one who is as ignorant as they are of lawyers' ways of thought; [this] is the great advantage to . . . judgment by [one's] peers." DEVLIN, *supra* note 2, at 140. Or, in the words of Sir William Holdsworth, "the jury system . . . tends to make the law intelligible by keeping it in touch with the common facts of life." *Id.* at 148.

[68] 1 ALEXIS DE TOCQUEVILLE, DEMOCRACY IN AMERICA 304 (1835).

> The jury is that portion of the nation to which the execution of the laws is entrusted, as the Houses of Parliament constitute that part of the nation which makes the laws; and in order that society may be governed with consistency and uniformity, the list of citizens qualified to serve on juries must increase and diminish with the list of electors.

Id; see also Carroll, *supra* note 57; Jocelyn Simonson, *The Criminal Court Audience in a Post-Trial World*, 127 HARV. L. REV. 2173, 2184–90 (2014); Herbert Mitgang, *'Inside the Jury Room'*, N.Y. TIMES, April 8, 1986, at 00017; *cf.* Jason Kreag, *The Jury's* Brady *Right*, 98 B.U. L. REV. 345 (2018) (arguing that the jury has a distinctively-legitimating role in the criminal justice system, even as distinct from the defendant's individual interests).

[69] The literature bemoaning our vanishing jury trial is voluminous. *See, e.g.*, Robert J. Conrad, Jr. & Katy L. Clements, *The Vanishing Criminal Jury Trial: From Trial Judges to Sentencing Judges*, 86 GEO. WASH. L. REV. 99 (2018) (presenting a judicial perspective of

for Devlin's thinking.[70] Even in his day, some 85% of criminal defendants chose disposition by a judge, and even among the remaining 15% slated for a jury, two-thirds would plead guilty.[71] So, out of every one hundred eligible prosecutions, only five were decided by jury. A far cry, then, from two centuries earlier when "[t]here was no plea bargaining in felony cases."[72]

The vanishing act has intensified with time. As John Langbein argued two decades ago (and the point has only become more pressing since),

---

the dramatic decline in federal criminal trials); NACDL, THE TRIAL PENALTY: THE SIXTH AMENDMENT RIGHT TO TRIAL ON THE VERGE OF EXTINCTION AND HOW TO SAVE IT (2018), available at www.nacdl.org/trialpenaltyreport [https://perma.cc/U9XE-3J7H] (presenting a defense perspective); Michael Vitiello, *Bargained-for-Justice: Lessons from the Italians?*, 48 U. PAC. L. REV 247 (2017) (analyzing Italy's hesitant but increasing reliance upon plea bargains); STEPHANOS BIBAS, THE MACHINERY OF CRIMINAL JUSTICE xvi–xvii (2012) (developing an argument that "[c]riminal justice used to be individualized, moral, transparent, and participatory but has become impersonal, amoral, hidden, and insulated from the people"); Jonathan A. Rapping, *Who's Guarding the Henhouse? How the American Prosecutor Came to Devour Those He Is Sworn to Protect*, 51 WASHBURN L. J. 513, 514 (2012) (examining the problem from the perspective of a public defender); Young, *supra* note 60 (presenting another judicial perspective); Ronald F. Wright, *Trial Distortion and the End of Innocence in Federal Criminal Justice*, 154 U. PA. L. REV. 79, 83 (2005) (developing a "trial distortion theory" of guilty pleas); GEORGE FISHER, PLEA BARGAINING'S TRIUMPH: A HISTORY OF PLEA BARGAINING IN AMERICA (2003) (meticulously chronicling the rise of plea bargaining in Massachusetts); Albert W. Alschuler, *Plea Bargaining and Its History*, 79 COLUM. L. REV. 1 (1979) (developing the history of plea bargaining); *see also* United States v. Stevenson, 2018 U.S. Dist. LEXIS 61988 at *1 (S.D.W.Va. Apr. 12, 2018) (explaining the rejection of multiple plea agreements as "not in the public interest[,] . . . transfer[ing] . . . criminal adjudications from the public arena to the prosecutor's office for the purpose of expediency at the price of confidence in and effectiveness of the criminal justice system"); Emily Yoffe, *Innocence Is Irrelevant*, ATLANTIC, Sept. 2017 ("By accepting the criminalization of everything, the bloat of the criminal-justice system, and the rise of the plea bargain, the country has guaranteed that millions of citizens will not have a fair shot at leading ordinary lives."); *cf.* Andrew Manuel Crespo, *The Hidden Law of Plea Bargaining*, 118 COLUM. L. REV. 1303, 1305–06 (2018) (arguing that scholars have failed to appreciate the "subconstitutional procedural law" regulating plea bargaining). The 'vanishing trial' literature is vast—on Westlaw, simple variants of the term appear in several hundred secondary sources. Thus, we highlight only a few pieces that will lead the interested reader to many more.

[70] *See* DEVLIN, *supra* note 2, at 129–33. Earlier studies in America date to the 1920s and 30s. FISHER, *supra* note 69, at 6–8. Thus it was that in 1928 Raymond Moley published "The Vanishing Jury." Raymond Moley, *The Vanishing Jury*, 2 S. CAL. L. REV. 97 (1928).

[71] DEVLIN, *supra* note 2, at 129–30, 176–77 n.1. The result was that "at the present day there are approximately five or six thousand trials by jury every year." *Id.* at 130.

[72] J. M. BEATTIE, CRIME AND THE COURTS IN ENGLAND, 1660-1800 at 336–37 (1986). Even if a bit overstated (*see, e.g.*, Alschuler & Deiss, *supra* note 59, at 922–23), and even though there is of course a critical difference between accepting a guilty plea and plea *bargaining*, the numbers are telling.

while the Sixth Amendment commands that "[i]n *all* criminal prosecutions, the accused shall enjoy the right to a speedy and public trial, by an impartial jury,"[73]

> a more accurate term to describe the use of jury trial in the discharge of our criminal caseload would be 'virtually none.' Like those magnificent guarantees of human rights that grace the pretended constitutions of totalitarian states, our guarantee of routine criminal jury trial is a fraud . . . . Can you find a hippopotamus in the Bronx? Yes, there's one in the Bronx Zoo, but it has nothing to do with life in the Bronx. It's a goner. And so, too, stunningly, is criminal jury trial.[74]

Lest we seem to wax too poetic, we fully recognize that the criminal jury trial of our founding was a "summary proceeding," incredibly wanting to modern sentiment.[75] We hardly desire a return to the oracular days in which criminal trials were hardly more reliable than a dice-roll, and it makes good sense that a world of much more information—including far better forensics—would lead to more negotiated settlements.[76]

---

[73] U.S. CONST. amend. VI (emphasis added). *See* AKHIL REED AMAR, AMERICA'S CONSTITUTION: A BIOGRAPHY 236 (2005) ("A criminal judge sitting without a criminal jury was simply not a duly constituted federal court capable of trying cases, just as the Senate sitting without the House was not a duly constituted federal legislature capable of enacting statutes.").

[74] Langbein, *supra* note 66, at 120–21. In the words of the Supreme Court in *Missouri v. Frye*, quoting scholars Robert Scott and Bill Stuntz, "plea bargaining . . . is not some adjunct to the criminal justice system; it *is* the criminal justice system." 566 U.S. 134, 144 (2012). In the words of George Fisher, plea bargaining has "triumphed," "driv[ing] our vanquished jury into small pockets of resistance." FISHER, *supra* note 69, at 1. And in the words of Laura Appleman, "We the People" have been "exile[d] from the justice system." LAURA I. APPLEMAN, DEFENDING THE JURY: CRIME, COMMUNITY, AND THE CONSTITUTION 3 (2015). Appleman continues, "Th[e] virtual elimination of the criminal jury trial has resulted in a system in which justice is dispensed by courts, bureaucrats, and prosecutors, with little room for the community's voice." *Id.* Similarly, Stephanos Bibas has questioned how "the criminal justice system bec[a]me so far removed from The People, who are nominally in charge." BIBAS, *supra* note 69, at xvi, xvii.

[75] *See* Langbein, *supra* note 66, at 122–23; *see also* John H. Langbein, *Understanding the Short History of Plea Bargaining*, 13 LAW & SOC'Y REV. 261 (1979) (further developing this thesis); John H. Langbein, *Torture and Plea Bargaining*, 46 U. CHI. L. REV. 3 (1978) (same).

[76] *See* LAWRENCE M. FRIEDMAN & ROBERT V. PERCIVAL, THE ROOTS OF JUSTICE: CRIME AND PUNISHMENT IN ALAMEDA COUNTY, CALIFORNIA, 1870-1910 (1981). In the view of Friedman and Percival,

> Perhaps there never was a golden age of trials. Except for really big cases, English trials in the eighteenth and nineteenth centuries were . . . [swift.] At any rate, the rise of the police and full-time prosecutors changed the conditions of criminal justice; so did "police science"–fingerprinting, blood tests, and so on. In a system run by amateurs, or lawyers who spent little bits of their time and energy, with no technology of detection or proof, a trial was perhaps as good a way as any to strain the guilty from the innocent. During the nineteenth century, however, criminal justice shifted

In other words, we are not trying to suggest that the decline of jury trials is only and straightforwardly negative, nor that, in theory, engaged judges could not provide the desirable role-reversible ethic just as elected legislators who craft our laws.[77] We leave a deeper dive into such matters for future work. Here our claim is more modest but fundamental nonetheless: one reason for preserving the criminal jury is because it is robustly role-reversible.

CONCLUSION

Despite many decades of improvement, artificial intelligence in the 'ideal sense'—embodied in broadly intelligent and capable 'thinking machines'—remains elusive, perhaps many decades away. Nevertheless, most experts are convinced this future is coming, and if it does, traditional dynamics of decision-making will come under strain.

But in crisis, opportunity: the dawn of powerful AI provides a fresh window into our democratic traditions, allowing us to better distinguish those worthy of preservation and to ask which traditions, despite their familiarity, have fallen short in practice. In this spirit, we have identified role-reversible judgment as one facet—alongside deliberation, voting, and other traditional indicators—of democratic legitimacy. Role-reversibility situates a decision-maker to appreciate the fortunes that come by grace or fate, enabling her to internalize the effects of judgment. This posture may improve outcomes. But even when it does not, it is central to our ideals of democratic self-authorship and is, therefore, of intrinsic value.

A 'but for grace' orientation thereby shifts the focus from humanity, as such, to reversibility. The relevant question is not who, or what kind of being, is entrusted to render judgment; it is whether they (or it) is reciprocally subject to judgment in reverse. The upshot is not that intelligent robots could never participate in our functions of judging. It is that to do so, they would have to be judgeable themselves. Which is to say,

---

> away from amateurs and part-timers toward full-time crime handlers. Once this change took place, it could no longer be assumed that trial by jury (in the idealized sense) was the normal way to handle a criminal case. After all, police and prosecutors had already "tried" the defendant. Why not leave it to them?

Id. at 194.

[77] Criminal defendants might question the 'invertibility' of a judge much more readily than that of a juror: 'Yeah, sure, you grow up how I grew up and see whether they make you a judge!' Again, we leave to future work whether a jury is necessarily essential to a democratically-legitimate criminal process, but there is no doubt it is more role-reversible.

they would have to be genuine equals, not overlords or underlings—just as any fellow citizen ought to be.