AI in Judicial Decision-Making*

Ignacio Cofone**

I. Introduction

Artificial intelligence (AI) is a slippery term. Throughout history, it has been used for different purposes and to identify different technologies. Broadly speaking, people (including computer scientists, the public, and the media) have used AI at each moment in time as an umbrella term that includes any major new technology that has capabilities usually identified only with humans. For the second half of the 20th century, the term was predominantly used to refer to robotics. In the last few decades, the umbrella term AI is used most frequently to refer to machine learning algorithms.[1] Having evolved from knowledge-based algorithms, machine learning algorithms are called AI because they have the ability to process enormous amounts of data and learn by example; something that, previously, only humans could do, and which revolutionizes the way in which we process data. Machine learning thus produces enormous social consequences. In the rest of the chapter, I will use the terms AI algorithms and machine learning algorithms interchangably.

AI algorithms—machine learning algorithms—have come in the last few decades to be used for novel purposes, among them judicial decision-making. Even if many

* Originally published in Florian Martin-Bariteau & Teresa Scassa, eds., *Artificial Intelligence and the Law in Canada* (LexisNexis, 2021).

** Assistant Professor and Norton Rose Fulbright Faculty Scholar, McGill University Faculty of Law. ignacio.cofone@mcgill.ca. The author is grateful to Stefanie Carsley, Rebecca Crootof, Katarina Daniels, and Thomas Kadri for their helpful comments. The author also thanks Malaya Powers and Ana Qarri for their outstanding research assistance and gratefully acknowledges the support of the Canadian Institute for the Administration of Justice Charles D. Gonthier Fellowship.

[1] Ignacio N. Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21 Stan. Tech. L. Rev. 167 at 179–180 (these include supervised and unsupervised learning, deep learning, reinforced learning, among others).

judges do not yet use AI, they are increasingly being presented with opportunities to do so. The introduction of AI algorithms in the courtroom is accompanied with both benefits and risks. The purpose of this chapter is to describe the way that AI is used and can be used in judicial decision-making, what risks this use entails, and how judges could address or control these risks.

AI algorithms are prediction machines. For example, they approach the question "is this loan candidate likely to be a good borrower?" by comparing the candidate's characteristics with the characteristics of other individuals who have and have not paid their debts. Similarly, they approach "is this person likely to flee if given parole?" by comparing the individual's characteristics with those of parolees who did or did not flee while on parole. More specifically, these algorithms extract the rules that they use to predict outcomes from their training data.[2] While rules-based algorithms (non-AI) are built on instructions indicating how to process data, machine learning algorithms are given large amounts of data for the algorithm to self-adjust.[3] Their key characteristic is that the model can learn its own decision rules to predict or estimate something.[4]

For the purposes of law reform, which concerns itself with identifying the aspects of AI that are disruptive to the law,[5] this rule extraction is the fundamental difference between AI and non-AI systems. This is because the rule-extraction characteristic of AI introduces emergence, meaning that what the AI will say or do is more unpredictable than rules-based systems. This rule-extraction characteristic therefore reduces foreseeability, which refers to how well each human involved can predict the outcome and therefore be deemed legally responsible for its consequences.[6]

Some AI have gained popularity in adjudication. COMPAS, for example, is a software owned by Equivant used by courts in the US to determine criminal defendants' likelihood of general and violent recidivism. TrueAllele is a software owned by Cybergenetics used to assess incomplete DNA samples—such as when the sample is small or there is DNA from more than one person in the sample. Different facial recognition softwares are used by police departments around the

---

[2] Karen Hao, "What is Machine Learning?" (17 November 2018), online: *M.I.T. Technology Review* https://www.technologyreview.com/s/612437/what-is-machine-learning-we-drew-you-another-flowchart/.

[3] David Lehr & Paul Ohm, "Playing with Data: What Legal Scholars Should Learn about Machine Learning" (2017) 51 U.C. Davis L. Rev. 653 at 671 (There are many ways to do this, such as decision tree learning, reinforcement learning, clustering, and Bayesian networks).

[4] David Lehr & Paul Ohm, "Playing with Data: What Legal Scholars Should Learn about Machine Learning" (2017) 51 U.C. Davis L. Rev. 653 at 672.

[5] Jack M. Balkin, "The Path of Robotics Law" (2015) 6 Cal. L. Rev. 45 at 50; Rebecca Crootof & BJ Ard, "Structuring Techlaw" (2021) 34 Harv. J. L. & Tech. [forthcoming].

[6] Ignacio N. Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21 Stan. Tech. L. Rev. 167 at 172, 183–186, 197.

world to identify suspects prior to charges and sentencing. The Government of Canada has considered using an AI (Tax Foresight) that determines the likelihood of someone being considered a tax avoider, which could inform litigation and audit decisions.[7] Similarly, a number of AI softwares are being proposed for Canadian courts that aim to weigh contradicting evidence[8] and predict trial outcomes.[9]

In what remains of the chapter, I first discuss in section 2 why AI is likely to be increasingly used by judges but unlikely to replace judges. Then, in section 3 I address how AI is and can be used by judges in substituting tasks, together with the error of trusting AI too little or too much. In section 4, I outline two key issues arising from the use of AI by judges: algorithmic bias being more difficult to detect and resolve than human bias, together with its consequences for accuracy and discrimination (4.1), and how to address algorithmic opacity so that using AI does not impede procedural and substantive rights (4.2). In section 5, I briefly identify some gaps in Canadian law related to those risks.

II. AI JUDGES AND JUDGES WHO USE AI

It seems unlikely that AI will replace judges, given how we understand law and adjudication. AI is, however, likely to increasingly assist judges in their decision-making. Understanding the issues that come up in this assistive role has consequences for how judges can best use AI.

It is unlikely that AI could replace judges altogether because adjudication is not a single, automatable activity. Machine learning works by extracting patterns in data; adjudication, on the other hand, is conducted through reasoning by analogy.[10] Machine learning cannot engage in analogical reasoning because analogies are not about extracting patterns, but rather about finding and interpreting common normative principles.[11] Relatedly, a key element of adjudication is the use of discretion, which machine learning cannot exercise because discretion crucially depends on context.[12]

---

[7] Dean Beeby, "Litigation Gone Digital: Ottawa Experiments with Artificial Intelligence in Tax Cases" (13 September 2018), online: *CBC News* https://www.cbc.ca/news/politics/artificial-intelligence-tax-justice-pilot-1.4817242.

[8] Wallis Snowdon, "Robot Judges? Edmonton Research Crafting Artificial Intelligence for Courts" (19 September 2017), online: *CBC News* https://www.cbc.ca/news/canada/edmonton/legal-artificial-intelligence-alberta-japan-1.4296763.

[9] See, *e.g.*, services offered by Loom Analytics at https://www.loomanalytics.com; Blue J. Legal at https://www.bluejlegal.com/ca; or SmartSettle at https://www.smartsettle.com.

[10] Cass R. Sunstein, "Of Artificial Intelligence and Legal Reasoning" (2001) 8 U. Chicago L. Sch. Roundtable 29 at 29, 31.

[11] Cass R. Sunstein, "Of Artificial Intelligence and Legal Reasoning" (2001) 8 U. Chicago L. Sch. Roundtable 29 at 33–34.

[12] Frank Pasquale, "A Rule of Persons, Not Machines: the Limits of Legal Automation" (2019) 87:1 Geo. Wash. L. Rev. 1 at 52–55; Richard M. Re & Alicia Solow-Niederman,

In judicial decision-making, like in administrative law, "[t]he domains in which machine learning might be of concern, at least as a *prima facie* matter, will be those in which artificial intelligence is used more for determining, rather than just supporting, decisions."[13] However, even relying on AI to determine some decisions does not mean it takes over the adjudicatory function. Adjudication is a uniquely human process because it involves several activities that change depending on context;[14] and the ways in which machines and humans process information and reach conclusions are relevantly different.[15] Judges make decisions based on legal and cultural knowledge, experience, their interactions with the expertise of others (lawyers, court-appointed experts, witnesses, etc.), context, and sometimes common sense, which machines (at least for now) cannot emulate.[16] Moreover, social and cultural norms can change the law rather abruptly in a way that would be difficult for a machine to identify, more difficult to react to, and impossible to advance itself—resulting in judgments that would not advance society and may not even reflect it. If AI were to take over an adjudication process, it would therefore change not only the mechanics of adjudication but also the adjudicatory values that are held by legal actors and that underlie the legal system.[17]

Legal norms in themselves (before examining their application) have been shown to be difficult to translate into code.[18] How would one automate, for example, legal standards such as the reasonable person standard? For AI to be useful in applying a

---

"Developing Artificially Intelligent Justice" (2019) 22 Stan. Tech. L. Rev. 242 at 252–253.

[13] Cary Coglianese & David Lehr, "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era" (2017) 105 Geo. L.J. 1147 at 1170.

[14] John Morison & Adam Harkens, "Re-engineering justice? Robot Judges, Computerized Courts and (semi) Automated Legal Decision-Making" (2019) 39:4 L. S. 618.

[15] Ian Kerr & Carissima Mathen, "Chief Justice John Roberts is a Robot" (2014) University of Ottawa Working Paper, online: https://ssrn.com/abstract=3395885; Rebecca Crootof, "'Cyborg Justice' and the Risk of Technological-Legal Lock-In" (2019) 119 Columbia L. Rev. Forum 233 at 236–242.

[16] Tania Sourdin, "Judge v Robot? Artificial Intelligence and Judicial Decision-Making" (2018) 41:4 U.N.S.W.L J. 1114 at 1128–1129; Quentin L. Koop, "Replacing Judges with Computers Is Risky" (20 February 2018), online: *Harvard Law Review Blog* https://blog.harvardlawreview.org/replacing-judges-with-computers-is-risky/; Frank Pasquale, "A Rule of Persons, Not Machines: the Limits of Legal Automation" (2019) 87:1 Geo. Wash. L. Rev. 1 at 29–30.

[17] Richard M. Re & Alicia Solow-Niederman, "Developing Artificially Intelligent Justice" (2019) 22 Stan. Tech. L. Rev. 242 at 244, 246–248, 252, 255; Frank Pasquale & Glyn Cashwell, "Prediction, Persuasion, and the Jurisprudence of Behaviourism" (2018) 68:Supp. 1 U. T. L. J. 63.

[18] See *e.g.* Lisa A. Shay, Woodrow Hartzog, John Nelson & Gregory Conti, "Do Robots Dream of Electric Laws? An Experiment in the Law as Algorithm" in Ryan Calo, A. Michael Froomkin & Ian Kerr, eds., *Robot Law* (Cheltenham, U.K.: Edward Elgar Publishing, 2016) 274.

rule, the rule must be non-vague;[19] but there is abundant philosophy of law literature showing that legal rules are almost always vague.[20] As judges know, interpreting and applying sources of law involves discretionary decisions. Judges, moreover, do more than adjudicate: they manage cases, aid in settling disputes, and develop and clarify legal norms. The same can be said about lawyers. Routine work—such as discovery, coding documents, and basic research functions—has been progressively automated throughout history,[21] but the complex tasks involved in most advice-giving will remain challenging to automate for the foreseeable future.[22]

Given these limitations, what will increasingly continue to happen is judges using AI to assist them in the adjudication process. Decisions such as assessing an offender's risk of recidivism,[23] determining which precedent has the most similar fact pattern,[24] or examining whether an incomplete DNA sequence found at the crime scene is likely to belong to the accused,[25] are likely to be increasingly delegated to AI. So are routine procedures, such as routine motion practice.[26]

III. SUBSTITUTING TASKS AND AUTOMATION BIAS

The integration of AI in a human-driven process—that is, "using" the AI—is sometimes referred to as "human in the loop." Under a human in the loop process,

---

[19] Harry Surden, "Artificial Intelligence and Law: An Overview" (2019) 35:4 Ga. St. U. L. Rev. 1305 at 1323; Rebecca Crootof, "'Cyborg Justice' and the Risk of Technological-Legal Lock-In" (2019) 119:7 Col. L. Rev. 233 at 239–240; Frank Pasquale, "A Rule of Persons, Not Machines: the Limits of Legal Automation" (2019) 87:1 Geo. Wash. L. Rev. 1 at 18–22.

[20] See *e.g.* Timothy Endicott, "Law is Necessarily Vague" (2001) 7:4 Leg. Theory 379; Hrafn Asgeirsson, *The Nature and Value of Vagueness in the Law* (Oxford, U.K.: Hart Publishing, 2020). From that point of view, AI may be more useful for textualist approaches than for other approaches. See Betsy Cooper, "Judges in Jeopardy!: Could IBM's Watson Beat Courts at Their Own Game?" (2011) 121:1 Yale L.J. 87.

[21] Carl Benedikt Frey & Michael A. Osborne, "The Future of Employment: How Susceptible are Jobs to Computerisation?" (2017) 114 Technol. Forecast. Soc. Change 254; Benjamin Alarie, Albert H. Yoon & Anthony Niblett, "How Artificial Intelligence Will Affect the Practice of Law" (2018) 68:Supp. 1 U.T.L.J. 106.

[22] Frank Pasquale & Glyn Cashwell, "Four Futures of Legal Automation" (2015) 63 U.C.L.A. L. Rev. Disc. 28 at 33–36, 39–45; Dana Remus & Frank Levy, "Can Robots be Lawyers?: Computers, Lawyers, and the Practice of Law" (2017) 30:3 Geo. J. Leg. Ethics 501.

[23] See *e.g. State v. Loomis* (2016) 371 Wis. (2d) 235 (Wis. Sup. Ct.).

[24] Dean Beeby, "Litigation Gone Digital: Ottawa Experiments with Artificial Intelligence in Tax Cases" (13 September 2018), online: *CBC News* https://www.cbc.ca/news/politics/artificial-intelligence-tax-justice-pilot-1.4817242.

[25] See *e.g. People v. Chubbs*, 2015 W.L. 139069 (Cal. 2nd Cir.).

[26] Tim Wu, "Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems" (2019) 119:7 Colum. L. Rev. 2001 at 2004–2005, 2021–2026.

a human has influence or oversight over the technology or, after the technology makes a prediction, the human can choose to follow or override the technology's recommended decision:

> Influence takes different forms. Sometimes, the human role is largely procedural: for example, pushing a given case up or back in the relevant queue, or deciding which cases merit more institutional resources. Other times, the human role is more dispositive, involving the power to shape outcomes, either in terms of a case's concrete effects (e.g., granting or denying benefits), or in terms of how the outcome is justified, or both. The specifics of the human role may vary, but the key is that a human has some form of meaningful discretion in particular cases.[27]

The COMPAS software, for example, makes a risk prediction but, ultimately, it is the role of the judge to decide whether to follow it. Thinking about the human in the loop means reframing the adoption of technology away from human replacement and towards designing and implementing technologies that are useful to human decision-makers, in this case adjudicators.

The challenge for judges will be to navigate the extent of their role in the human in the loop process: determining to what extent it is appropriate to rely on the AI. Judges will need to know when and how they should follow the advice provided by an AI. Using AI in this assistive manner can be useful for increasing efficiency and uniformity in the legal system.[28] AI is faster, cheaper, and, for some tasks, more accurate than human decision-makers.[29] However, for this to be effective, the right amount of deference (not more, not less) must be given to the machine. And decision-makers often defer too much.

If a judge does not trust the outcome of an AI system, the system becomes unhelpful—algorithmic predictions are of little use if routinely ignored. The tendency to place too little reliance on an AI after seeing it err is called algorithmic aversion.[30] The tendency to place too much reliance on its outcomes, however, may

---

[27] Kiel Brennan-Marquez, Karen Levy & Daniel Susser, "Strange Loops: Apparent versus Actual Human Involvement in Automated Decision Making" (2019) 34:3 B.T.L.J. 745 at 749.

[28] Richard M. Re & Alicia Solow-Niederman, "Developing Artificially Intelligent Justice" (2019) 22:2 Stan. Tech. L. Rev. 242 at 246, 255; Anthony Niblett, Benjamin Alarie & Albert H. Yoon, "Regulation by Machine" (2016), online: *SSRN* https://ssrn.com/abstract=2878950.

[29] Vasant Dhar, "When to Trust Robots with Decisions, and When Not To" (17 May 2016), online: *Harvard Business Review* https://hbr.org/2016/05/when-to-trust-robots-with-decisions-and-when-not-to; Benjamin Alarie, "The Path of the Law: Towards Legal Singularity" (2016) 66:4 U.T.L.J. 443 at 449–450.

[30] Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err" (2015) 144:1 J. Exp. Psychol. Gen. 114 at 119–126; Berkeley J. Dietvorst, Joseph P. Simmons & Cade Massey, "Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms if They Can (Even Slightly) Modify Them" (2016) 64:3 Manag. Sci. 1155; Berkeley Dietvorst & Soaham Bharti,

be worse. If a judge relies on the algorithm's conclusion too much, without weighing it against other evidence, her experience, her cultural and legal knowledge, her interaction with other experts, and context, *inter alia*, incorporating the AI may actually deteriorate the adjudication process. The tendency to overtrust an AI is called automation bias.[31] Automation bias is concerned with how human decision-makers, including judges, treat algorithms—unlike algorithmic biases, which are biases of the algorithm itself. To some extent, "[t]o effectively participate in a human–machine team, the person in the loop must have an appropriately calibrated amount of trust."[32] While algorithmic aversion can certainly occur when the errors of an algorithm become evident to decision-makers, decision-makers show a marked tendency towards automation bias, sometimes with dire outcomes.[33]

People's automation bias blurs, to some degree, the distinction between using AI and being replaced by it. This falls under what Jack Balkin calls the substitution effect. "[T]hrough their interactions with robots and AI systems, people are willing to substitute them for animals or human beings *in certain contexts and for certain purposes*"[34] by delegating to them (often mechanical) tasks that they no longer want to do directly. This delegation is always partial because a task is being delegated, not a person's role.[35] People are never completely replaced by AI, but they sometimes use AI to perform some of the tasks that they themselves, or others, previously performed. The specific risks of using AI, Balkin argues, stem from the inadvertent replacement of humans with machines in specific tasks.[36]

While I have claimed that AI is more likely to assist human judges than to replace them, a reader may object by pointing to instances of what appears to be AI replacing some relatively minor adjudicative functions. For example, in Estonia, AI

"People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error" (2020) Psychol. Sci. [forthcoming], online: *SSRN* https://ssrn.com/abstract=3424158.

[31] See *e.g.* Kate Goddard, Abdul Roudsari & Jeremy C. Wyatt, "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators" (2012) 19:1 J. Am. Med. Inform. Assoc. 121.

[32] Rebecca Crootof, "'Cyborg Justice' and the Risk of Technological-Legal Lock-In" (2019) 119 Columbia L. Rev. Forum 233 at 243.

[33] Rebecca Crootof, "'Cyborg Justice' and the Risk of Technological-Legal Lock-In" (2019) 119 Columbia L. Rev. Forum 233 at 244; Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" (2019) 5 Engaging Sci. Technol. Soc. 40 at 41–42.

[34] Jack M. Balkin, "The Path of Robotics Law" (2015) 6 Cal. L. Rev. Circuit 45 at 57 [emphasis added].

[35] Jack M. Balkin, "The Path of Robotics Law" (2015) 6 Cal. L. Rev. Circuit 45 at 57–59.

[36] Jack M. Balkin, "2016 Sidley Austin Distinguished Lecture on Big Data Law and Policy: The Three Laws of Robotics in the Age of Big Data" (2017) 78:5 Ohio State L. J. 1217 at 14–16.

algorithms are used in small claims courts to arbitrate cases. In the same vein, DoNotPay offers claimants a "robot lawyer" chatbot that provides minimal legal advice, such as helping claimants with parking ticket appeals and a service called "Do Not Sign" that analyzes license agreements.[37] A reader may ask then, "are these not examples of AI replacing humans?" The answer is no, because even if AI replaces the full set of tasks that are now done by any specific person, it can never replace the adjudicator. The human adjudicator will just be required to play a different role in the decision-making process: decision-making moves from the person performing those tasks directly to the person designing, selecting, applying, and overseeing the technology. There is always a human making the decisions, even when the human decides to neglect her role.[38] In other words, because of the nature of the activity that the law regulates, there is always a human in the loop. The issue at stake is how the social relationship between humans, in this case between decision-makers and citizens, changes when intermediated by this technology.

To the extent that judges use AI in this assistive manner, two key risks arise in the administration of justice that they can address: algorithmic bias and opacity.

IV. RISKS AND OPPORTUNITIES

1. Inaccurate or Discriminatory Rulings Due to Algorithmic Bias

All human beings—even the most well-meaning—have implicit biases that can surface when making decisions. Algorithms, instead, are sometimes presented as fair and unbiased decision-making agents. But it is well documented that algorithmic impartiality is a myth: there exists a plethora of examples and documented cases in which decision-making algorithms have produced biased outcomes.[39] As AI increasingly assists judges, understanding algorithmic bias is important because it can lead to inaccurate and discriminatory results.

The promise of fair and unbiased algorithmic decisions is a myth not because algorithms are not useful to help humans make decisions (they are). It is a myth

---

[37] See DoNotPay at https://donotpay.com/.See also Shannon Liao, "'World's first robot lawyer' now available in all 50 states" (12 July 2017), online: *The Verge* https://www.theverge.com/2017/7/12/15960080/chatbot-ai-legal-donotpay-us-uk; Jon Porter, "This 'robot lawyer' can take the mystery out of license agreement" (20 November 2019), online: *The Verge* https://www.theverge.com/2019/11/20/20973830/robot-lawyer-donotpay-ai-startup-license-agreements-sign-arbitration-clauses.

[38] John Nay & Katherine J. Strandburg, "Generalizability: Machine Learning and Humans-in-the-Loop" in Roland Vogl, ed., *Research Handbook on Big Data Law* (Edward Elgar, 2020), online: *SSRN* https://ssrn.com/abstract=3417436.

[39] See *e.g.* Julia Angwin & Jeff Larson, "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say" (30 December 2016), online: *ProPublica* www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say; Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias Against Women" (9 October 2018), online: *Reuters* https://reut.rs/2Od9fPr.

because algorithmic bias is as real as it is inevitable—and it must be regulated differently than human bias.[40] There are three types of bias in AI that can lead to inaccurate and discriminatory results: bias in the process of building the algorithmic model, bias in the sample that is used to train the algorithm, and societal biases captured and amplified by the algorithm.

The first type of bias is a biased process. This is a bias in how an algorithm processes information.[41] Biases in an algorithmic process often exist because human biases are translated into the system.[42] Even if no human chooses the outcome directly, there is always human involvement in how that outcome is arrived at: humans frame the problem and make a choice about what the algorithm should predict before any data are processed. Once that is decided, there is human involvement through gathering data to train the algorithm and selecting the variables that the algorithm should consider (the *features*). There are always, at some level, human decision-makers that influence the process.

The second type is a biased sample data. An algorithm's predictive power is only as good as the data that it is fed. If an algorithm mines in a section of a dataset that, for any reason, is unrepresentative of the population, it will produce non-representative outputs (i.e. inaccurate and potentially discriminatory individual decisions).[43] Individual records, for example, may suffer from quality problems due to partial or incorrect data. The entire dataset might also have quality problems at higher rates for an entire protected group compared to others or might be unrepresentative of the general population.[44]

The third type of algorithmic bias is data that reflects societal biases. A machine learning algorithm's training data may reflect prior systemic discrimination.[45] An AI can thus produce a disparate impact or indirect discrimination even when correctly trained with representative data.[46] The difference between biased sample data and

---

[40] Kate Crawford & Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms" (2014) 55:1 Boston College L. Rev. 93 at 124–128.

[41] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L. J. 1389 at 1399–1402.

[42] Batya Friedman and Helen Nissenbaum, "Bias in Computer Systems" (1996) 14:3 A.C.M. Trans. Inf. Syst. 330, at 333–336 (explaining the difference between preexisting bias and technical bias).

[43] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L.J. 1389 at 1402–1404.

[44] Solon Barocas & Andrew D. Selbst, "Big Data's Disparate Impact" (2016) 104:3 Cal. L. Rev. 671 at 680–681, 684–687.

[45] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L.J. 1389 at 1404–1406.

[46] Solon Barocas & Andrew D. Selbst, "Big Data's Disparate Impact" (2016) 104:3 Cal. L. Rev. 671 at 673–674, 691.

this type of bias is that, here, the data is representative of the population, but this representative data still produces a disparate outcome because of embedded social inequalities.[47]

Examples of these biases exist in practically every area of decision-making where AI is used, but perhaps the most widely known among them is the use of the COMPAS software for risk assessment in criminal procedure. COMPAS aims to predict the likelihood that an accused will recidivate if granted parole.[48] A few years ago, a ProPublica investigation accused COMPAS of producing racially biased results for both high-risk and low-risk classifications.[49] The investigation found that almost twice as many black defendants as white defendants were incorrectly classified as high-risk by COMPAS, and white defendants were also more likely to be incorrectly classified as low-risk than were black defendants.[50] COMPAS is still widely used by judges in the US for parole hearings, bail hearings, and sometimes sentencing.

Other risk assessment instruments employed in criminal justice exhibit similar patterns of bias in their outcomes. For example, a tool used at the federal level in Canada to make probation decisions (Post Conviction Risk Assessment) was found to give black offenders higher average post-conviction risk assessment scores than white offenders.[51] Using criminal history as a predictor captures societal biases (the

---

[47] Aylin Caliskan, Joanna J. Bryson & Arvind Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human-Like Biases" (2017) 356:6334 Science 183; Daniel Rosenberg, "Data Before Fact" in Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron* (Cambridge, Mass.: M.I.T. Press, 2013) 15.

[48] Timm Brennan, William Dieterich & Beate Ehret, "Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System" (2009) 36:1 Crim. Justice Behav. 21 at 22–24.

[49] Julia Angwin, Jeff Larson, Surya Mattu & Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks" (23 May 2016), online: *ProPublica* www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Jeff Larsen, Surya Mattu, Lauren Kirchner & Julia Angwin, "How We Analyzed the COMPAS Recidivism Algorithm" (23 May 2016), online: *ProPublica* www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm.

[50] See Alexandra Chouldechova, "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments" (2016) 5:2 Big Data 153 at 153, 156; Anupam Chander, "The Racist Algorithm?" (2017) 115:6 Mich. L. Rev. 1023 at 1033; Fiona Doherty, "Obey All Laws and Be Good: Probation and the Meaning of Recidivism" (2016) 104:2 Geo. L. J. 291 at 352–353; Jessica M. Eaglin, "Constructing Recidivism Risk" (2017) 67 Emory L.J. 59 at 96; Melissa Hamilton, "Risk-Needs Assessment: Constitutional and Ethical Challenges" (2015) 52 Am. Crim. L. Rev. 231 at 239–241; Kelly Hannah-Moffat, "Algorithmic Risk Governance: Big Data Analytics, Race and Information Activism in Criminal Justice Debates" (2019) 23:4 Theor. Criminol. 453 at 461.

[51] Jennifer L. Skeem & Christopher T. Lowenkamp, "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact" (2016) 54:4 Criminology 680 at 685–700.

third type of algorithmic bias mentioned above) because criminal history captures the relationship between race and arrest, where black individuals are more likely to be arrested than white individuals for the same level of criminal activity.[52]

If left unchecked, these biases can lead judges to inadvertently endorse inaccurate or discriminatory outcomes in their rulings.

2. Impaired Procedural Rights Due to Opacity and Private Interests

Opacity refers to individuals' lack of knowledge or understanding of how an algorithm arrived at its output (the decision) from its input (the data that it is fed). Understanding algorithmic opacity is important for defining the scope of individuals' right to an explanation in algorithmic decision-making and the scope of private interests in keeping AI systems secret when they are used in the administration of justice.

Opacity is a problem for AI in judicial decision-making because AI systems are biased in different ways than the humans whose functions they may replace, and opacity makes those algorithmic biases more difficult to detect and reduce.[53] Opacity also interferes with the right to an explanation that people subject to the decisions of these machines may have.[54] Opacity complicates the application of doctrines that protect people's civil rights, such as the doctrine of indirect discrimination,[55] because it is difficult to correct a decision-making process that one can neither access nor understand.[56] Disclosing the algorithmic processes used in judicial decision-making, for that reason, is important for protecting people's civil rights, promoting accountability, and guaranteeing procedural fairness.[57]

Opacity can be classified into three types: intentional opacity, literacy-driven opacity, and inherent opacity.[58] Intentional opacity occurs when the process through which an AI reaches its result is deliberately hidden. Usually, this is a result of

---

[52] Jennifer L. Skeem & Christopher T. Lowenkamp, "Risk, Race, and Recidivism: Predictive Bias and Disparate Impact" (2016) 54:4 Criminology 680 at 700.

[53] Miriam C. Buiten, "Towards Intelligent Regulation of Artificial Intelligence" (2019) 10:1 Eur. J. Risk Regul. 41 at 43.

[54] Margot E. Kaminski, "The Right to Explanation, Explained" (2019) 34:1 B.T.L.J. 189.

[55] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L. J. 1389 at 1427–1433.

[56] Danielle Keats Citron & Frank Pasquale, "The Scored Society: Due Process for Automated Predictions" (2014) 89:1 Wash. L. Rev. 1 at 18–20.

[57] See *e.g.* Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System" (2018) 70:5 Stan. L. Rev. 1343; Sonia K. Katyal, "Private Accountability in the Age of Artificial Intelligence" (2019) 66:1 U.C.L.A. L. Rev. 54 at 120.

[58] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms" (2016) 3:1 Big Data Soc. 1 at 1–2.

gaming concerns or concerns about competitors' unfair advantages.[59] For example, COMPAS is protected by trade secrets, so its code, training data, and model are unavailable to the public and to those who are granted or refused bail or parole based on its results. Literacy-driven opacity refers to the fact that the general public does not understand how AI processes work because it would be too costly to teach. Inherent opacity refers to a certain degree of opacity that is a natural and inevitable result of the characteristics of some algorithms.[60] For example, deep learning algorithms, which have less human intervention during the process, make it more difficult to scrutinize how outcomes are reached.[61]

Each type of opacity demands a different response from judges and lawmakers. Intentional opacity can be mitigated by simply mandating disclosure of the decision-making process, which in turn helps reveal whether antidiscrimination law is being disregarded.[62] Literacy-driven opacity can be redressed by implementing expert auditing and engaging the question of for whom the algorithm should be made transparent. Judges in jury trials, for example, can address literacy-driven opacity by ensuring that jurors have enough tools at their disposal to interpret an algorithm's results.[63] Finally, because the law has experience addressing human decision-makers as black boxes, one can derive lessons from how the law has responded to human opacity in exploring responses to inherent algorithmic opacity. For instance, we require ex-post explanations from many human decision-makers through verbal accounts of their decision process, such as arguments in judicial rulings, and we could demand equivalent accounts for AI—requiring ex-post explanations, unlike requiring algorithmic interpretability, is possible with a black box system.[64]

These considerations impact how one should conceive of a right to explanation for AI in judicial decision-making, where individuals have a right to understand how a decision was made. The decision about how to create a socially beneficial right to explanation must pay close attention to the type of judicial decision involved, the impact that each type of explanation will have on individuals, and the social benefit created by each explanation.

---

[59] Ignacio N. Cofone & Katherine J. Strandburg, "Strategic Games and Algorithmic Secrecy" (2019) 64:4 McGill L.J. 621 at 626–632.

[60] Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms" (2016) Big Data Soc. 1 at 1–2.

[61] See Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter & Lalana Kagal, "Explaining Explanations: An Overview of Interpretability of Machine Learning" (3 February 2019), online: *arXiv* https://arxiv.org/pdf/1806.00069.pdf.

[62] See generally Rebecca Wexler, "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System" (2018) 70:5 Stan. L. Rev. 1343.

[63] Andrea Roth, "Trial by Machine" (2016) 104:5 Geo. L.J. 1245 at 1296–1297.

[64] See Zachary C. Lipton, "The Mythos of Model Interpretability" (2018) 16:3 A.C.M. Queue, online: *A.C.M. Queue* https://queue.acm.org/detail.cfm?id=3241340.

The most common objections to recognizing a right to explanation are concerns about trade secrecy and citizens "gaming the system" when presented with a right to explanation, both of which refer to intentional opacity.[65] Trade secrecy privilege is sometimes invoked to avoid disclosing any or all of the elements of an algorithmic decision.[66] This privilege, particularly when used in criminal law to withhold information from defendants, has been accused of being ahistorical, harmful, and unnecessary.[67] Judges should weigh the competing interests of protecting private actors' trade secrets and the public interest. They can push back against these private interests and mandate disclosure as appropriate. For example, judges can request filing under seal to limit the information being distributed beyond the trial and beyond the needs of the proceeding, or issue protective orders that prevent using or distributing the information protected by trade secrecy.[68] These measures can compatibilize private interests with the benefits of transparency.

Judges should, in turn, be critical of and interrogate gaming claims by private and public actors in the courtroom. Gaming concerns are frequently overstated. When people change their behaviour to obtain more favourable algorithmic outcomes, the result is not always socially undesirable.[69] People's strategic behaviour can be socially beneficial if it makes them more truly deserving of a beneficial decision or if it corrects errors caused by an algorithm's predictive inaccuracy. The consequential over-secrecy deprives society not only of the benefits of disclosure to individuals, but also of improvements in decision quality that could result from disclosure, improving accountability. Judges' view of the public interest may thus weigh in favour of mandating disclosure notwithstanding gaming claims.

Useful disclosure of information about a machine learning algorithm, more importantly, could take many forms—it is not an either-or matter.[70] Depending on

---

[65] Inherently opaque deep learning models are rare for the types of decision-making algorithms that are useful to judges. And even with these models, some disclosure is always possible; for example, the algorithm's code, output variables, and training data.

[66] See *e.g. State v. Loomis*, 371 Wis. (2d) 235 (Wis. Sup. Ct. 2016); *People v. Chubbs*, 2015 W.L. 139069 (Cal. 2nd Cir.).

[67] Rebecca Wexler, "Life, Liberty, and Trade Secrets" (2018) 70:5 Stan. L. Rev. 1343 at 1395–1429.

[68] Rebecca Wexler, "Life, Liberty, and Trade Secrets" (2018) 70:5 Stan. L. Rev. 1343 at 1409–1413.

[69] Ignacio N. Cofone & Katherine J. Strandburg, "Strategic Games and Algorithmic Secrecy" (2019) 64:4 McGill L.J. 621 at 624–625.

[70] Andrew D. Selbst & Solon Barocas, "The Intuitive Appeal of Explainable Machines" (2018) 87:3 Fordham L. Rev. 1085; Sandra Wachter, Brent Mittelstadt & Chris Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR" (2018) 31:2 Harv. J. L. & Tech. 841; Sandra Wachter & Brent Mittelstadt, "A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI"(2019) 2 Colum. Bus. L. Rev. 494.

concerns about trade secrets and the possibility of individuals gaming the system, disclosures aimed at the right to explanation could include the data used to train an algorithm, the code, the features, the feature weights, the model, or the output variables.[71]

The usefulness of each of these types of disclosures will vary among different groups of people. For an average person, the most useful information will be the variables that the model takes into account. Experts may find the code or the model most useful to audit the process. The code, which is the type of disclosure that usually faces the greatest resistance to being disclosed, is mostly useful for algorithmic auditing.[72] Disclosure will therefore have different policy goals depending on context. Oftentimes, features will be most useful for accountability, given that features tend to be more easily understandable than the formulas for combining them. For instance, in the COMPAS example, what defendants may find most useful is knowing the features (*e.g.*, charges, criminal history, substance abuse, education, work), as auditing the model or the code requires technical expertise that they may not have. Decision-makers can disclose the features, or the data that is used to assemble the features, without disclosing the model that determines the weight given to each feature in the overall assessment and without disclosing the code. But journalists or associations with more resources than individual defendants may find the model or the code useful to identify whether there are systemic effects like the ones ProPublica found for COMPAS. In those cases, even disclosing training data sources alone is often helpful for accountability, for example, to facilitate evaluating whether the learned model is biased towards a group of individuals.

Because effective gaming requires fairly extensive information about the features employed and how they are combined by the algorithm, it is always possible to disclose some relevant information about the decision—for example, what features are used and how they are combined—without creating a significant gaming threat. Identifying not only whether gaming is a plausible concern, but also what information for that AI facilitates gaming allows, at a minimum, for selective disclosure of other types of information.[73] Banks and credit bureaus, for example, seem to understand this idea, as credit bureaus disclose the features used to make credit score assessments and banks have open mortgage risk-assessment metrics.

Identifying the appropriate scope of private secrecy interests and the alternatives for selective disclosure are important for developing a right to explanation for AI in judicial decision-making. Legal discussions on algorithmic transparency in the

---

[71] Ignacio N. Cofone & Katherine J. Strandburg, "When Does Gaming Justify Algorithmic Secrecy?" (2020) *NYU Law and Economics Research Paper*.

[72] See *e.g. People v. Chubbs*, 2015 W.L. 139069 (Cal. 2nd Cir. 2015); *State v. Loomis*, 371 Wis. (2d) 235 (Wis. Sup. Ct. 2016).

[73] Ignacio N. Cofone & Katherine J. Strandburg, "Strategic Games and Algorithmic Secrecy" (2019) 64:4 McGill L.J. 621 at 653–654.

administration of justice should thus focus less on *whether* to require disclosure and more on *what* information can be disclosed.[74]

V. KEY GAPS IN THE LAW

1. Preventing Algorithmic Bias Through Data Governance

It is as dangerous as it is inaccurate to hold that, because algorithmic decision-making is computational, it cannot discriminate. As with human decision-makers, algorithmic decision-making can produce discriminatory outcomes even without discriminatory intent.[75] Legal scholars have shown how AI algorithms can lead to discriminatory outcomes based on the use of people's personal information.[76]

Antidiscrimination law is ill-equipped to handle these biases in two ways. First, antidiscrimination law and other retroactive legal remedies do not deal effectively with systemic discrimination,[77] which is what AI can capture, perpetuate, and amplify. Second, while an AI algorithm might not directly consider protected categories, it can pick up on biases in the sample data and existing societal or historical biases and rely on proxies related to protected categories, leading to indirect discrimination that is difficult to detect and resolve. The type of bias that traces most clearly to antidiscrimination law categories is a biased process, as treating two groups differently maps onto direct discrimination if one of those groups is a protected category.[78] But sample bias and the amplification of social biases are difficult to classify into the category of indirect discrimination.

Algorithmic discrimination, therefore, must be regulated differently from human discrimination.[79] Algorithmic biases may be addressed through a preventive approach by privacy rules.[80] Since they regulate the collection, storage, and

---

[74] Ignacio N. Cofone & Katherine J. Strandburg, "Strategic Games and Algorithmic Secrecy" (2019) 64:4 McGill L.J. 621 at 625.

[75] See *e.g.* Julia Angwin & Jeff Larson, "Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say" (30 December 2016), online: *ProPublica* https://www.propublica.org/article/bias-in-criminal-risk-scores-is-mathematically-inevitable-researchers-say; Jeffrey Dastin, "Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women" (9 October 2018), online: *Reuters* https://reut.rs/2Od9fPr.

[76] See *e.g.* Pauline T. Kim, "Data-Driven Discrimination at Work" (2017) 58:3 Wm. & Mary L. Rev. 857 at 869–892.

[77] See Colleen Sheppard, "Institutional Inequality and the Dynamics of Courage" (2013) 31:2 Windsor Y.B. Access Just. 103.

[78] Solon Barocas & Andrew D. Selbst, "Big Data's Disparate Impact" (2016) 104 Cal. L. Rev. 671.

[79] Kate Crawford & Jason Schultz, "Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms" (2014) 55:1 Boston College L. Rev. 93 at 124–128.

[80] Ignacio N. Cofone, "Antidiscriminatory Privacy" (2019) 72:1 S.M.U. L. Rev. 139 at 169–173.

dissemination of people's personal information, data protection statutes such as the *Personal Information and Electronic Documents Act*[81] are equipped to reduce algorithmic discrimination in a way that antidiscrimination law cannot: by regulating the personal information that algorithms are trained with.[82] Algorithmic discrimination that is problematic in terms of systemic or indirect discrimination stems from bias in the data that the algorithm is trained with. Therefore, data governance rules can reduce those biases by shaping such data.[83]

Because these gaps in the law exist, judges have a crucial role to play in helping to address them. But most importantly, their role as decision-makers is to understand these biases and their implications for accuracy and discrimination when deciding how much to trust the outcome of an AI in adjudication.

2. Equipping Judges with Tools to Address Opacity

Judges must have tools to curtail secrecy demands motivated by private interests and determine what form of the right to explanation should apply in each decision. This should depend on the type of decision, the impact on individuals, the level of concern with disclosure (trade secrecy or gaming), and the broader social benefit created by the explanation. There will be times when judges have discretion about whether to look behind the curtain and examine how a specific AI functions. In these circumstances, there are good reasons to favour transparency and there are methods to protect the private interests of companies that develop algorithms while obtaining the social benefits of transparency, such as protective orders.

While the three kinds of opacity discussed are specific to algorithms, the problem of opacity more generally is not unique to automated decision-making. Human decision-makers, in some sense, are also black boxes, as we can ask them for the reasons for their decisions but we cannot examine the decision-making process directly. The difference among them is that the law already treats humans as a black box, for example by asking them to explain their decisions, but does not yet do the same with algorithms, which it often mistakenly assumes to be impartial and unbiased.[84] The failure to scrutinize algorithmic opacity with the same vigour as human opacity in decision-making can have problematic consequences. For example, someone who wants to discriminate can take advantage of an algorithm's unscrutinized opacity, instead of making the decision themselves, to conceal

[81] S.C. 2000, c. 5.

[82] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L. J. 1389 at 1406–1427.

[83] Ignacio N. Cofone, "Algorithmic Discrimination is an Information Problem" (2019) 70:6 Hastings L. J. 1389 at 1406–1427.

[84] Katherine J. Strandburg, "Rulemaking and Inscrutable Automated Decision Tools" (2019) 119:7 Colum. L. Rev. 1851 at 1863–1864.

discriminatory intent.[85]

Furthermore, while human opacity may be inevitable—and the best we can do is demand good-faith explanations—algorithmic opacity is not, so the automation of some judicial decision-making tasks may be an opportunity to further expand, not just avoid reducing, citizens' civil rights. For that reason, if the opacity problem is adequately regulated, the progressive incorporation of AI in decision-making may, counterintuitively, place those who exert oversight, such as courts of appeal, in a better position for detecting bias, indirect discrimination, and other decision errors than they were before.

---

[85] Nicholas Diakopoulos, "Accountability in Algorithmic Decision Making" (2016) 59:2 Commun. A.C.M. 56 at 59–61.