NUS Law Working Paper 2020/011

Through A Glass, Darkly: Artificial Intelligence and The Problem of Opacity

Simon Chesterman

chesterman@nus.edu.sg

[April 2020]

THROUGH A GLASS, DARKLY: ARTIFICIAL INTELLIGENCE AND THE PROBLEM OF OPACITY

*Simon Chesterman**

As computer programs become more complex, the ability of non-specialists to understand how a given output has been reached diminishes. Opaqueness may also be built into programs to protect proprietary interests. Both types of system are capable of being explained, either through recourse to experts or an order to produce information. Another class of system may be naturally opaque, however, using deep learning methods that are impossible to explain in a manner that humans can comprehend. An emerging literature describes these phenomena or specific problems to which they give rise, notably the potential for bias against specific groups. Drawing on examples from the United States, the European Union, and China, this article develops a novel typology of three discrete regulatory challenges posed by opacity. First, it may encourage — or fail to discourage — inferior decisions by removing the potential for oversight and accountability. Secondly, it may allow impermissible decisions, notably those that explicitly or implicitly rely on protected categories such as gender or race in making a determination. Thirdly, it may render illegitimate decisions in which the process by which an answer is reached is as important as the answer itself. The means of addressing some or all of these concerns is routinely said to be through transparency. Yet while proprietary opacity can be dealt with by court order and complex opacity through recourse to experts, naturally opaque systems may require novel forms of 'explanation' or an acceptance that some machine-made decisions cannot be explained — or, in the alternative, that some decisions should not be made by machine at all.

---

* Dean and Provost's Chair Professor, National University of Singapore Faculty of Law. I am deeply grateful to Damian Chalmers, Miriam Goldby, Hu Ying, Arif Jamal, Jeong Woo Kim, Koh Kheng Lian, Lau Kwan Ho, Emma Leong, Lin Lin, Daniel Seng, Sharon Seah, David Tan, Tan Zhong Xing, Umakanth Varottil, anonymous reviewers, and others for their comments on earlier versions of this text. Invaluable research assistance was provided by Violet Huang, Eugene Lau, Ong Kye Jing, and Yap Jia Qing. Errors and omissions are due to the author alone.

INTRODUCTION

Eric Loomis was 31 when he was arrested in La Crosse, Wisconsin, in connection with a drive-by shooting. Two rounds from a sawn-off shotgun had been fired at a house a little after 2 a.m. on a Monday morning in February 2013. Though no one was injured, police were called and soon identified Loomis's Dodge Neon two miles away. A short car chase ended when he crashed into a snowdrift; together with a passenger he continued on foot, but was apprehended and charged with reckless endangerment and possession of a firearm. Loomis denied involvement in the shooting, pleading guilty to lesser charges of fleeing a police officer and driving a stolen vehicle.[1]

These were all repeat offences. Loomis was also a registered sex offender, stemming from an earlier conviction for sexual assault, and on probation for dealing in prescription drugs. His lawyer nevertheless argued for mitigation, highlighting a childhood spent in foster homes where he had been subjected to abuse; with an infant son of his own, Loomis was now training to be a tattoo artist. Prior to sentencing, the circuit court ordered a risk assessment using software known by the acronym COMPAS.[2] Based on information gathered from a defendant's criminal file and an interview, COMPAS generates scores on a scale from one to ten, indicating the predicted likelihood that he or she will commit further crimes.

Equivant,[3] the company that developed COMPAS, regards the proprietary algorithm that generates these scores as a trade secret. The scores themselves are not. Neither Loomis nor his lawyer was able to see or to question how the figures had been reached, but the presiding judge cited them in justifying a six-year prison sentence. 'You're identified,' Judge Scott Horne said, 'through the COMPAS assessment, as an individual who is at high risk to the community.' The judge then ruled out probation 'because of the seriousness of the crime and because your history, your history on supervision, and the risk assessment tools that have been utilized, suggest that you're extremely high risk to re-offend.'[4]

Opacity is the antithesis of legal decisions. Accountability for those decisions typically requires that the decision-maker has a convincing reason

---

[1] *2 Arrested in La Crosse Drive-by Shooting*, NEWS8000.COM, Feb. 11, 2013; Anne Jungen, *Driver Gets 8½ Years in Drive-by Shooting, Drug Case*, LA CROSSE TRIBUNE, Aug. 13, 2013; Mitch Smith, *In Wisconsin, a Backlash Against Using Data to Foretell Defendants' Futures*, N.Y. TIMES, June 22, 2016.

[2] COMPAS stands for Correctional Offender Management Profiling for Alternative Sanctions.

[3] The company was formerly known as Northpointe, Inc.

[4] *State v. Loomis*, 881 N.W.2d 749, 755 (Wis., 2016).

for a decision or act. Judicial decisions in particular give special weight to reasoning.[5] In the common law tradition, only the *ratio decidendi* — the legal basis for the decision — is binding on lower courts. Appeals to higher courts look for errors in the law or in its application to the facts as disclosed in the reasons. The failure to give reasons can itself be a ground of appeal in its own right. [6] Eric Loomis's sentencing decision appeared to violate these principles. The judge's reliance on COMPAS was criticized by academics and civil society, and was central to an appeal that made its way — almost — to the U.S. Supreme Court.[7]

The problem of understanding artificial intelligence (A.I.) systems is not new. [8] In *The Black Box Society*, Frank Pasquale compared the role of algorithms in the modern world to Plato's metaphor of the cave, with the general public trapped and able only to see 'flickering shadows cast by a fire behind them'; the prisoners cannot comprehend the actions, let alone the agenda, of those who create the images that are all they know of reality.[9] More prosaically, it has been argued that computer simulation displaces humans from the center of the epistemological enterprise. For most of human history, the expansion of knowledge meant the expansion of human knowledge and understanding. The emergence of computational methods that transcend our abilities presents what Paul Humphreys calls the 'anthropocentric predicament'. [10] Distinct from the challenges posed by

[5] Herbert Wechsler, *Toward Neutral Principles of Constitutional Law*, 73 HARV. L. REV. 1, 19-20 (1959) (arguing that the 'virtue or demerit of a judgment turns … entirely on the reasons that support it').

[6] There are, of course, exceptions to this. Juries, for example, are not required to give reasons for the limited decisions they make within the legal system. *See generally* Mathilde Cohen, *When Judges Have Reasons Not to Give Reasons: A Comparative Law Approach*, 72 WASH. & LEE L. REV. 483 (2015).

[7] *See* section III.B, *infra*.

[8] For a discussion of attempts to define A.I., *see* ARTIFICIAL INTELLIGENCE: A MODERN APPROACH 1-5 (Stuart J. Russell and Peter Norvig eds., 3rd ed. 2010). Four broad approaches can be identified: acting humanly (the famous Turing test), thinking humanly (modelling cognitive behavior), thinking rationally (building on the logicist tradition), and acting rationally (a rational-agent approach favored by Russell and Norvig as it is not dependent on a specific understanding of human cognition or an exhaustive model of what constitutes rational thought). Though much of the literature focuses on 'general' or 'strong' A.I. (meaning the creation of a system that is capable of performing any intellectual task that a human could) the focus in this article is on the more immediate challenges raised by 'narrow' A.I. — meaning systems that can apply cognitive functions to specific tasks typically undertaken by a human.

[9] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 190 (2015).

[10] Paul Humphreys, *The Philosophical Novelty of Computer Simulation Methods*, 169 SYNTHESE 615, 617 (2009).

autonomy in A.I. systems,[11] the increasing opacity of those systems is not a challenge to the centrality of human agents as legal *actors* so much as a challenge to our ability to understand and evaluate *actions* — something essential to meaningful regulation.[12]

'Opacity' is used here to mean the quality of being difficult to understand or explain. As in the case of COMPAS, this may be due to certain technologies being proprietary. To protect an investment, detailed knowledge of the inner workings of a system may be limited to those who own it. A second form of opacity may arise from complex systems that require specialist skills to understand them. Such systems often evolve over time, being added to by different stakeholders, but are in principle capable of being explained.[13]

Neither of these forms of opacity — proprietary or complex — pose particularly new problems for law. Intellectual property law has long recognized protection of intangible creations of the human mind and exceptions based on fair use.[14] To deal with complex issues, governments and judges routinely have recourse to experts.[15] The same cannot be said of a third reason for opacity, which is systems that are naturally opaque. Some deep learning methods are opaque effectively by design, as they rely on reaching decisions through machine learning rather than, for example, following a decision tree that would be transparent, even if it might be complex.[16]

---

[11] *See* Simon Chesterman, *Artificial Intelligence and the Problem of Autonomy*, 1 NOTRE DAME J. EMERGING TECH. (forthcoming).

[12] The term 'regulation' is chosen cautiously. Depending on context, its meaning can range from any form of behavioral control, whatever the origin, to the specific rules adopted by government that are subsidiary to legislation. BARRY M. MITNICK, THE POLITICAL ECONOMY OF REGULATION: CREATING, DESIGNING, AND REMOVING REGULATORY FORMS (1980); ANTHONY OGUS, REGULATION: LEGAL FORM AND ECONOMIC THEORY (2004); THE OXFORD HANDBOOK OF REGULATION (Robert Baldwin, Martin Cave, and Martin Lodge eds., 2010); TONY PROSSER, THE REGULATORY ENTERPRISE: GOVERNMENT, REGULATION, AND LEGITIMACY 1-6 (2010). For present purposes, the focus is on public control of a set of activities.

[13] A paradox is that availability of large amounts of information may give the illusion of transparency. *See, e.g.,* Cynthia Stohl, Michael Stohl, and Paul M. Leonardi, *Managing Opacity: Information Visibility and the Paradox of Transparency in the Digital Age*, 10 INT'L J. COMM. 123 (2016) (distinguishing between visibility and transparency).

[14] Amanda Levendowski, *How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem*, 93 WASH. L. REV. 579 (2018).

[15] *See, e.g.,* CAROL A.G. JONES, EXPERT WITNESSES: SCIENCE, MEDICINE, AND THE PRACTICE OF LAW (1994).

[16] Machine learning denotes the ability of a computer to improve on its performance without being specifically programmed to do so. This process may be supervised or unsupervised, or through a process of reinforcement: KEVIN P. MURPHY, MACHINE LEARNING: A PROBABILISTIC PERSPECTIVE 2 (2012). *Cf.* Jenna Burrell, *How the Machine*

To pick a trivial example, the programmers of Google's AlphaGo could not explain how it came up with the strategies for the ancient game of *Go* that defeated the human grandmaster, Lee Sodol, in 2016. Lee himself later said that in their first game the program made a move that no human would have played — and which was only later shown to have planted the seeds of its victory.[17]

Such output-based legitimacy — optimal ends justifying uncertain means — is appropriate in some areas. Medical science, for example, progresses based on the success or failure of clinical trials with robust statistical analysis. If the net impact is positive, the fact that it may be unclear precisely *how* a procedure or pharmaceutical achieves those positive outcomes is not regarded as a barrier to allowing it into the market.[18] Though patient autonomy means that important decisions are made by the individual most affected, tolerance for adverse effects is built into the process, with patients advised as to the risks of negative as well as positive outcomes.[19]

Legal decisions, on the other hand, are generally not regarded as appropriate for statistical modelling. Though certain decisions may be expressed in terms of burdens of proof — balance of probabilities, beyond reasonable doubt, and so on — these are to be determined in individualized assessments of a given case, rather than based on a forecast of the most likely outcomes from a larger set of cases.[20]

"Thinks": Understanding Opacity in Machine Learning Algorithms, 3 BIG DATA & SOC'Y 1 (2016). 'Decision tree' is used here in the sense of a static set of parameters specified in advance and to be applied consistently. This is distinct from decision tree models that are themselves developed through machine learning.

[17] *Google's A.I. Beats World Go Champion in First of Five Matches*, BBC NEWS, Mar. 9, 2016. A subsequent version, AlphaGo Zero, was taught only the rules of Go and in three days had mastered the ancient game. In match ups against the version that beat the human grandmaster, Lee Sodol, the newer version beat the old 100 to zero. *See* David Silver et al., *Mastering the Game of Go Without Human Knowledge*, 550 NATURE 354 (10/18/online 2017).

[18] Alex John London and Jonathan Kimmelman, *Why Clinical Translation Cannot Succeed Without Failure*, 4 ELIFE e12844 (2015); Riccardo Miotto et al., *Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records*, 6 SCIENTIFIC REPORTS 26094 (05/17/online 2016), *at* http://https://doi.org/10.1038/srep26094. Research into mental illness in particular is fraught with uncertainty as to the underlying causes of disease and the mechanisms that bring about cures. *See* ANNE HARRINGTON, MIND FIXERS: PSYCHIATRY'S TROUBLED SEARCH FOR THE BIOLOGY OF MENTAL ILLNESS (2019).

[19] Patients are, of course, provided with individualized assessment based on their condition, history, and so on. But the use of objective population-based trends is generally accepted. Omer Gottesman et al., *Guidelines for Reinforcement Learning in Healthcare*, 25 NATURE MEDICINE 16 (2019).

[20] On the impact of A.I. on the legal profession more generally, see KEVIN D. ASHLEY, MODELING LEGAL ARGUMENT: REASONING WITH CASES AND HYPOTHETICALS (1990);

There is a growing literature criticizing reliance on algorithmic decision-making with legal consequences. A significant portion now focuses on opacity,[21] highlighting specific concerns such as bias,[22] or seeking remedies

---

PETER WAHLGREN, AUTOMATION OF LEGAL REASONING: A STUDY ON ARTIFICIAL INTELLIGENCE AND LAW (1992); GIOVANNI SARTOR, ARTIFICIAL INTELLIGENCE AND LAW (1993); RICHARD SUSSKIND, THE FUTURE OF LAW: FACING THE CHALLENGES OF INFORMATION TECHNOLOGY (1996); RICHARD SUSSKIND, TRANSFORMING THE LAW: ESSAYS ON TECHNOLOGY, JUSTICE, AND THE LEGAL MARKETPLACE (2000); RICHARD SUSSKIND, THE END OF LAWYERS? RETHINKING THE NATURE OF LEGAL SERVICES (2008); DORY REILING, TECHNOLOGY FOR JUSTICE: HOW INFORMATION TECHNOLOGY CAN SUPPORT JUDICIAL REFORM (2010); RICHARD SUSSKIND, TOMORROW'S LAWYERS: AN INTRODUCTION TO YOUR FUTURE (2013); RICHARD SUSSKIND AND DANIEL SUSSKIND, THE FUTURE OF THE PROFESSIONS: HOW TECHNOLOGY WILL TRANSFORM THE WORK OF HUMAN EXPERTS (2015); JOANNA GOODMAN, ROBOTS IN LAW: HOW ARTIFICIAL INTELLIGENCE IS TRANSFORMING LEGAL SERVICES (2016); KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE (2017); RICHARD SUSSKIND, ONLINE COURTS AND THE FUTURE OF JUSTICE (2019).

[21] *See, e.g.,* Burrell, *supra* note 16; Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. PA. L. REV. 633 (2017); Jane Bambauer and Tal Zarsky, *The Algorithm Game*, 94 NOTRE DAME L. REV. 1 (2018); Seth Katsuya Endo, *Technological Opacity & Procedural Injustice*, 59 B.C. L. REV. 821 (2018); Sandra Wachter, Brent Mittelstadt, and Chris Russell, *Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J.L. & TECH. 841 (2018); Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 STAN. L. REV. 1343 (2018); Karen Yeung, *Algorithmic Regulation: A Critical Interrogation*, 12 REG. & GOVERNANCE 505 (2018); Dan L. Burk, *Algorithmic Fair Use*, 86 U. CHI. L. REV. 283 (2019); Michael E. Donohue, *A Replacement for Justitia's Scales?: Machine Learning's Role in Sentencing*, 32 HARV. J.L. & TECH. 657 (2019); Kirsten Martin, *Ethical Implications and Accountability of Algorithms*, 160.J. BUS. ETHICS 835 (2019); Katherine J. Strandburg, *Rulemaking and Inscrutable Automated Decision Tools*, 119 COLUM. L. REV. 1851 (2019); Leah Wisser, *Andora's Algorithmic Black Box: The Challenges of Using Algorithmic Risk Assessments in Sentencing*, 56 AM. CRIM. L. REV. 1811 (2019); Ronald Yu and Gabriele Spina Alì, *What's Inside the Black Box? A.I. Challenges for Lawyers and Researchers*, 19 LEGAL INFORMATION MANAGEMENT 2 (2019); Monika Zalnieriute, Lyria Bennett Moses, and George Williams, *The Rule of Law and Automation of Government Decision-Making*, 82 MOD. L. REV. 425 (2019).

[22] *See, e.g.,* Oscar H. Gandy, Jr., *Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems*, 12 ETHICS AND INFORMATION TECHNOLOGY 29 (2010); Solon Barocas and Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CAL. L. REV. 671 (2016); Anupam Chander, *The Racist Algorithm?*, 115 MICH. L. REV. 1023 (2017); Sharad Goel et al., *Combatting Police Discrimination in the Age of Big Data*, 20 NEW CRIM. L. REV. 181 (2017); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 WM. & MARY L. REV. 857 (2017); Levendowski, *supra* note 14; Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein, *Discrimination in the Age of Algorithms*, 10 J. LEGAL ANALYSIS 113 (2018); James A. Allen, *The Color of Algorithms: An Analysis and Proposed Research Agenda for Deterring Algorithmic Redlining*, 46 FORDHAM URB. L.J. 219 (2019); Richard Berk, *Accuracy and Fairness for Juvenile Justice Risk Assessments*, 16 J. EMPIRICAL LEGAL STUD. 175 (2019); Ignacio N. Cofone, *Algorithmic Discrimination Is an Information Problem*, 70 HASTINGS L.J. 1389

through transparency.[23] Yet the challenges of opacity go beyond bias and will not all be solved through calls for transparency or 'explainability'. Drawing on well-known examples and arguments from the United States and the European Union, as well as less-studied innovations in China, this article develops a novel typology of those challenge posed by proprietary, complex, and natural opacity. The first is that 'black box' decision-making may lead to inferior decisions. Accountability and oversight are not merely tools to punish bad behavior; they also encourage good behavior. Excluding that possibility reduces opportunities to identify wrongdoing, as well as the chances that decisions will be subjected to meaningful scrutiny and thereby be improved. Secondly, opaque decision-making practices may provide cover for impermissible decisions, such as through masking or reifying discrimination. Even if statistical models suggested that persons of a particular race should be given longer prison sentences, for example, acting on such predictions would not be tolerated in a judge and should not be accepted in an A.I. system. Finally, the legitimacy of certain decisions depends on the transparency of the decision-making process as much as on the decision itself. Judicial decisions are the best, but not the only, example of this.

These challenges reflect discrete reasons for wariness of opaque decisions. The quality of outcomes approaches the question through a utilitarian lens and a desire for better decisions. The avoidance of impermissible decisions reflects deontic concerns — decisions that should

(2019); Talia B. Gillis and Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019); Aziz Z. Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 DUKE L.J. 1043 (2019); Sandra G. Mayson, *Bias In, Bias Out*, 128 YALE L.J. 2218 (2019); Sarah Valentine, *Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control*, 46 FORDHAM URB. L.J. 364 (2019); Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. (forthcoming).

[23] *See, e.g.,* Kiel Brennan-Marquez, *"Plausible Cause": Explanatory Requirements in the Age of Powerful Machines*, 70 VAND. L. REV. 1249 (2017); Philipp Hacker and Bilyana Petkova, *Reining in the Big Promise of Big Data: Transparency, Inequality, and New Regulatory Frontiers*, 15 NW. J. TECH. & INTELL. PROP. 1 (2017); Kroll et al., *supra* note 21; Mike Ananny and Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOCIETY 973 (2018); Robert Brauneis and Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103 (2018); Paul B. de Laat, *Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?*, 31 PHIL. & TECH. 525 (2018); Andrew Selbst and Solon Barocas, *The Intuitive Appeal of Explainable Machines*, 87 FORDHAM L. REV. 1085 (2018); Bryan Casey, Ashkan Farhangi, and Roland Vogl, *Rethinking Explainable Machines: The GDPR's "Right to Explanation" Debate and the Rise of Algorithmic Audits in Enterprise*, 34 BERKELEY TECH. L.J. 145 (2019); Vincent Chiao, *Fairness, Accountability, and Transparency: Notes on Algorithmic Decision-Making in Criminal Justice*, 15 INT'L J.L. IN CONTEXT 126 (2019); Ashley Deeks, *The Judicial Demand for Explainable Artificial Intelligence*, 119 COLUM. L. REV. 1829 (2019).

not be allowed even if they are in some sense 'optimal'.[24] Legitimacy, by contrast, relies upon proper authority and process — with authority derived not so much from the quality of the decision as from the publicness of the reasoning.

The means of addressing some or all of these concerns is routinely said to be through transparency. Yet while proprietary opacity can be dealt with by court order and complex opacity through recourse to experts, naturally opaque systems may require novel forms of 'explanation' or an acceptance that some machine-made decisions cannot be explained — or, in the alternative, that some decisions should not be made by machine at all.

I. INFERIOR DECISIONS

Technology can be made opaque to protect an investment but also to prevent scrutiny. Such scrutiny may reveal trade secrets or it may reveal incompetence. At its most venal, opaqueness provides cover for the intentional manipulation of outcomes or to thwart investigation. Volkswagen, for example, wrote code that gamed tests used by regulators to give the false impression that vehicle emissions were lower than in normal usage.[25] Uber similarly designed a version of its app that identified users whose behavior suggested that they were working for regulators in order to limit their ability to gather evidence.[26]

A more general problem is that even good faith inscrutability may prevent interrogations of data quality. In some cases, greater transparency might reveal how much data is being used, giving rise to privacy concerns.[27] In others, the patchiness of data might be revealed, raising questions about the reliability of the process or the confidence level of the outcome.[28] This

[24] Roger Brownsword and Alon Harel, *Law, Liberty, and Technology: Criminal Justice in the Context of Smart Machines*, 15 INT'L J.L. IN CONTEXT 107, 112 (2019).

[25] EPA, California Notify Volkswagen of Clean Air Act Violations/Carmaker Allegedly Used Software that Circumvents Emissions Testing for Certain Air Pollutants (U.S. Environmental Protection Agency, Washington, DC, Sept. 18, 2015), *at* http://19january2017snapshot.epa.gov/newsreleases/epa-california-notify-volkswagen-clean-air-act-violations-carmaker-allegedly-used_.html.

[26] Leslie Hook, *Uber Used Fake App to Confuse Regulators and Rivals*, FINANCIAL TIMES, Mar. 4, 2017; Michael Guihot, Anne F. Matthew, and Nicolas P. Suzor, *Nudging Robots: Innovative Solutions to Regulate Artificial Intelligence*, 20 VAND. J. ENT. & TECH. L. 385, 426 (2017).

[27] Karl Manheim and Lyric Kaplan, *Artificial Intelligence: Risks to Privacy and Democracy*, 21 YALE J.L. & TECH. 106 (2019).

[28] *See, e.g.,* Chris Reed, *How Should We Regulate Artificial Intelligence?*, 376

phenomenon of 'garbage in, garbage out' is as old as the first computer. Charles Babbage, the English polymath who fashioned a mechanical device often credited as such, raised the issue in 1864. His memoir recalls twice being asked by members of Parliament whether putting wrong figures into his difference engine might nonetheless lead to the right answers coming out. 'I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question,' he observed.[29]

Human complacency and automation bias make these more than theoretical problems. As human involvement in a process — notionally 'in' or 'over' the loop[30] — is reduced to its most mechanistic, the tendency to accept default suggestions increases.[31] This may be compared with the danger posed by autonomous vehicles operating at a level where the human 'driver' may release the wheel — but is expected to remain ready to seize back control at any moment. In reality, humans are generally unable to maintain for any length of time the attention necessary serve the function of backup driver in an emergency; several car manufacturers have announced that they plan to skip this level of automation completely.[32] That is an example of complacency. Bias arises due to the tendency of most people to ascribe to an automated system greater trust in its analytical capabilities than in their own.[33]

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A: MATHEMATICAL, PHYSICAL AND ENGINEERING SCIENCES (2018) (discussing research into pneumonia that revealed errors, and the conclusion that neural nets ran the risk of embedding those errors in an undetectable manner that would increase patient risk); Sandra Wachter and Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and A.I.*, 2019(2) COLUM. BUS. L. REV. 1 (2019).

[29] CHARLES BABBAGE, PASSAGES FROM THE LIFE OF A PHILOSOPHER 67 (1864).

[30] A commonly used metaphor is of a human being in, over, or out of a decision-making process referred to as a 'loop'. 'Human-in-the-loop' refers to decision-making supported by the system, for example through suggesting options or recommendations, but with the human taking positive decisions. 'Human-over-the-loop' denotes a process in which the human can oversee the process and make interventions as necessary. 'Human-out-of-the-loop' means the process runs with minimal or no human intervention. Yeung, *supra* note 21, at 508.

[31] Steven P.R. Rose and Hilary Rose, *"Do not Adjust Your Mind, There Is a Fault in Reality" — Ideology in Neurobiology*, 2 COGNITION 479, 498-99 (1973). On the larger impact of anchoring in sentencing decisions, see Birte Enough and Thomas Mussweiler, *Sentencing Under Uncertainty: Anchoring Effects in the Courtroom*, 31(7) J. APPLIED SOC. PSYCHOL. 1535 (2001).

[32] Paresh Dave, *Google Ditched Autopilot Driving Feature After Test User Napped Behind Wheel*, REUTERS, Oct. 31, 2017; *Why Car-Makers Are Skipping Sae Level-3 Automation?*, M14 INTELLIGENCE, Feb. 20, 2018.

[33] Raja Parasuraman and Dietrich Manzey, *Complacency and Bias in Human Use of Automation: An Attentional Integration*, 52(3) HUMAN FACTORS 381, 392 (2010); Robert Challen et al., *Artificial Intelligence, Bias and Clinical Safety*, 28 B.M.J. QUALITY & SAFETY 231 (2019).

A related problem is that such systems may also provide cover for human agents. A survey of lawyers and judges in Canada, for example, found that many regarded software like COMPAS as an improvement over subjective judgment: though risk assessment tools were not deemed especially reliable predictors of future behavior, they were also favored because using them minimized the risk that the lawyers and judges themselves would be blamed for the consequences of their decisions.[34]

Addressing complacency and automation bias goes far beyond the regulatory challenges that are the focus of this article. For present purposes, it is sufficient to observe that they should not be a basis for avoiding accountability in the narrow sense of being obliged to give an account of a decision, even if after the fact, or to avoid responsibility for harm as a result of that decision.

As in many areas of technology regulation, the European Union offers comparatively stronger protections under its General Data Protection Regulation (GDPR),[35] which makes clear that the right not to be subject to automated processing cannot be avoided by 'token' human involvement. Routine acceptance of automated processes would not suffice; meaningful oversight requires a person with authority and competence to review a decision — including having access to 'all the relevant data'.[36] The limits of those protections will be discussed in section III.B, below.

The notion that opacity leads to inferior decisions has a long history in software development. Combined with a resistance to proprietary opacity, this insight lies at the heart of the open source movement.[37] Complete openness will not be appropriate or possible in all circumstances, but the idea that it should not be limited simply in order to prevent external scrutiny seems uncontroversial. Such questions are more challenging as the systems become more complex and the outputs less susceptible to objective evaluation.

---

[34] Kelly Hannah-Moffat, *The Uncertainties of Risk Assessment: Partiality, Transparency, and Just Decisions*, 27(4) FED. SENTENCING REP. 244 (2015).

[35] General Data Protection Regulation 2016/679 (GDPR) 2016 (EU), art 22.

[36] Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679 (Article 29 Data Protection Working Party, 17/EN WP251rev.01, Oct. 3, 2017), *at* http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=612053, at 20-21; Maja Brkan, *Do Algorithms Rule the World? Algorithmic Decision-Making and Data Protection in the Framework of the GDPR and Beyond*, 27 INT'L J.L. & INFO. TECH. 91, 101-02 (2019).

[37] Sheen S. Levine and Michael J. Prietula, *Open Collaboration for Innovation: Principles and Performance*, 25(5) ORG. SCI. 1287 (2014).

II. IMPERMISSIBLE DECISIONS

One of the benefits of automated decision-making is that it can reduce the arbitrariness of human decisions. Given a large number of similar questions, properly programmed computers will provide predictable and consistent answers. Whereas many evaluative decisions made by humans are based on unconscious group biases and intuitive reactions, algorithms follow the parameters set out for them.[38] They are only as good as the data they are given and the questions they are asked, however. In practice, algorithms can reify existing disparities — and, as we shall see, the absence of conscious bias in specific decisions may actually frustrate attempts to rectify those disparities by relying on anti-discrimination laws.[39]

A prominent example is screening decisions. Many industries now use A.I. systems to simplify repetitive processes such as reviewing job applications, assessing creditworthiness, setting insurance premiums, detecting fraud, and so on. These systems often rely on two discrete algorithms: the screening algorithm itself selects candidates from the pool or assigns them a score; this in turn may be based on a training algorithm, which uses data to improve the screening algorithm.[40]

Used well, screening processes efficiently and consistently treat like cases alike. This is most effective in binary decisions, such as whether an email is spam or whether a transaction is fraudulent. There is an objective answer using a predefined category — 'spam' or 'fraud' — with answers that are verifiable in a manner upon which most evaluators of that decision would agree. False positives and negatives can be flagged for the training algorithm, which feeds back to the screening algorithm and progressively reduces those errors.

Problems arise when more contested categories are invoked, such as fairness, or when such algorithms are used in order to predict future behavior by specific individuals,[41] such as how well they will perform in a particular job — or whether they will commit another crime. In some cases, the results are perverse. An audit of one résumé-screening algorithm identified that the

[38] Gandy, *supra* note 22, at 32; Goel et al., *supra* note 22. This may be particularly useful in decision-making systems that are delegated and distributed: Strandburg, *supra* note 21, at 1857.

[39] Barocas and Selbst, *supra* note 22; Karen Yeung, *Five Fears About Mass Predictive Personalization in an Age of Surveillance Capitalism*, 8(3) INT'L DATA PRIVACY L. 258 (2018); Brent Daniel Mittelstadt et al., *The Ethics of Algorithms: Mapping the Debate*, [2016] BIG DATA & SOC'Y, 7-8 (2016).

[40] Kleinberg et al., *supra* note 22.

[41] Chelsea Barabas et al., *Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment*, 81 PROC. MACHINE LEARNING RES. 1 (2018).

two most important factors indicative of job performance at a particular company were being named Jared and having played high school lacrosse.[42] In others, reliance upon algorithms may reflect or reify discriminatory practices.

A. How Bias Is Learned

Bias can be 'learned' in at least two ways. If overt prejudice affects the data used to train algorithms, that prejudice may be replicated.[43] But if an algorithm is used to draw inferences based on a sample population, it is also possible that unintended biases may be revealed due to the training data itself, the selection and weighting of variables, or the manner in which outputs are interpreted.[44] Various scholars compare this to the distinction between 'disparate treatment', or intentional behavior, and 'disparate impact' in U.S. civil rights jurisprudence.[45] An example of the former is Amazon's résumé-screening algorithm, which was trained on ten years of data but had to be shut down when programmers discovered that it had 'learned' that women's applications were to be regarded less favorably than men's.[46]

Examples of unintended bias would include facial recognition software that is less effective at recognizing dark-skinned faces because its training tends to be done using light-skinned ones.[47] The use of unrepresentative data is not unique to A.I. systems, of course. A meta-analysis of psychology studies found that the vast majority of those published relied on the participation of western university students, who were then treated as representative of all of humanity.[48] Different problems can arise with selection and weighting of variables. An ostensibly neutral metric like productivity of employees, for example, might adversely impact women if it does not account for the fact that they are more likely than men to take

---

[42] Dave Gershgorn, *Robot Indemnity: Companies Are on the Hook if Their Hiring Algorithms Are Biased*, QUARTZ, Oct. 22, 2018.

[43] Valentine, *supra* note 22.

[44] Selena Silva and Martin Kenney, Algorithms, Platforms, and Ethnic Bias: An Integrative Essay (University of California, Berkeley, BRIE Working Paper 2018-3, 2018).

[45] *Ricci v DeStefano*, 557 U.S. 557 (2009). *See, e.g.,* Barocas and Selbst, *supra* note 22, at 694-712; Kim, *supra* note 22, at 866; Zachary C. Lipton, Alexandra Chouldechova, and Julian McAuley, *Does Mitigating ML's Impact Disparity Require Treatment Disparity?*, ARXIV 1711.07076v3 (2018); Gillis and Spiess, *supra* note 22, at 461. For a wider discussion of benchmarks for algorithmic discrimination, see Huq, *supra* note 22, at 1115-23.

[46] Cofone, *supra* note 22, at 1397-98; Strandburg, *supra* note 21, at 1852.

[47] Manheim and Kaplan, *supra* note 27, at 159; Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. REV. 54, 69 (2019).

[48] Joseph Henrich, Steven J Heine, and Ara Norenzayan, *The Weirdest People in the World?*, 33(1) BEHAV. & BRAIN SCI. 61 (2010) (the title refers to subjects being drawn entirely from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies).

maternity leave.[49]

Perhaps the greatest risk comes with the interpretation of outputs, which brings us back to risk assessment tools like COMPAS. A widely cited report by *ProPublica* concluded in 2016 that COMPAS correctly predicted recidivism in nearly two-thirds of cases, but that its false positives and false negatives were both skewed against African Americans. Of those who did not reoffend, African Americans were almost twice as likely to have been labelled 'high risk' as compared with whites; of those who did go on to commit further crimes, whites were almost twice as likely to have been deemed 'low risk'.[50] The report was criticized for oversimplifying risk assessment, cherry-picking results, and ignoring the higher incarceration rates of African Americans.[51] It was also challenged on the basis that it failed to acknowledge that data-driven risk assessments have repeatedly been shown to be superior to professional human judgments, which themselves are prone to bias.[52]

These debates join a rich literature defending and critiquing the use of actuarial risk assessments in the United States, where standardized decision-making from the 1970s focused on prevention of future crime and has been linked with ongoing problems of mass incarceration generally, and the jailing of African American men in particular.[53] The emergence of proprietary and otherwise opaque tools like COMPAS has exacerbated the concerns about such models, due to complacency and automation bias, but the underlying problem is one of the oldest of logical fallacies: *cum hoc ergo propter hoc* (with this, therefore because of this). Or, as it is rendered in introductory texts on statistics: correlation does not imply causation.

Risk assessments originally used regression models. Regression in statistics is a tool that identifies a set of variables that are predictive of a given outcome. Model checking and selection enables the identification of optimal

---

[49] *Cf.* Rafael Lalive et al., *Parental Leave and Mothers' Careers: The Relative Importance of Job Protection and Cash Benefits*, 81(1) REV. ECON. STUD. 219 (2014).

[50] Julia Angwin et al., *Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks*, PROPUBLICA, May 23, 2016, *at* http://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[51] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp, *False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks"*, 80(2) FEDERAL PROBATION 38 (2016).

[52] Alexandra Chouldechova, *Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments*, 5(2) BIG DATA 153 (2017).

[53] Malcolm M. Feeley and Jonathan Simon, *The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications*, 30 CRIMINOLOGY 449 (1992); Paula Maurutto and Kelly Hannah-Moffat, *Assembling Risk and the Restructuring of Penal Control*, 46 BRIT. J. CRIMINOLOGY 438 (2006); Barabas et al., *supra* note 41.

weights for those variables that best predict the outcome of interest.[54] The COMPAS 'violent recidivism risk score', for example, is calculated through an equation that weighs history of violence and noncompliance against age, age at first arrest, and level of education. As the company's manual notes, it is similar to the way in which a car insurance company estimates the risk of a customer having an accident.[55] The algorithm's impenetrability, however, and the criticism to which that gave rise anticipate future challenges as A.I. systems become more complex and play a greater role in decisions affecting the rights and obligations of individuals.

Supervised machine learning techniques embody many of the problems of regression, in that the goal is prediction. Though some studies have shown that machine learning is more accurate than traditional statistical methods, this comes at the expense of transparency.[56] Here opacity becomes a concern as the black box nature of some of these techniques both obscures the decision-making process while also creating — in the minds of some users, at least — the illusion of greater sophistication and, therefore, reliability.

Scholars in the field continue to argue over the extent to which social, economic, and psychological factors need to be taken into account in improving the accuracy of risk assessment models.[57] A more fundamental challenge questions the purpose of using such models in the first place.

Risk assessments like COMPAS use historical data to predict future behavior. There are two basic objections to this. The first is that punishment should generally be meted out by the state only for crimes committed in the past rather than those that might be committed in the future. Though the prospects of reoffending might properly be considered when choosing from a range of possible sentences, or when considering early release, truly preventive detention is rare in most well-ordered jurisdictions.[58] The second objection is that the application of summary statistics to individuals is the

---

[54] ANDREW GELMAN AND JENNIFER HILL, DATA ANALYSIS USING REGRESSION AND MULTILEVEL/HIERARCHICAL MODELS (2007).

[55] A Practitioner's Guide to COMPAS Core (Northpointe, 2015), *at* http://assets.documentcloud.org/documents/2840784/Practitioner-s-Guide-to-COMPAS-Core.pdf, at 29.

[56] Grant Duwe and KiDeuk Kim, *Sacrificing Accuracy for Transparency in Recidivism Risk Assessment: The Impact of Classification Method on Predictive Performance*, 1(3) CORRECTIONS 155 (2016).

[57] *See, e.g.,* Kelly Hannah-Moffat, *Sacrosanct or Flawed: Risk, Accountability and Gender-responsive Penal Politics*, 22 CURRENT ISSUES IN CRIMINAL JUSTICE 193 (2011); Seth J. Prins and Adam Reich, *Can We Avoid Reductionism in Risk Reduction?*, 22(2) THEORETICAL CRIMINOLOGY 258 (2018). Mental health, for example, tends to be excluded in favor of more measurable and statistically significant covariates. Barabas et al., *supra* note 41, at 5.

[58] HALLIE LUDSIN, PREVENTIVE DETENTION AND THE DEMOCRATIC STATE (2016).

very definition of stereotyping.[59] The fact that a person comes from a community with higher rates of crime may make it more probable that he or she will commit a crime, but that is not a basis for punishing him or her for it in advance.[60]

Interesting parallels may be drawn here with the use of personally identifying data by police. To the extent that authorities rely on fingerprints and DNA samples collected from those who have been arrested or convicted in the past, it significantly increases the likelihood that these identifiers will be used against that group in the future, entrenching discriminatory practices.[61] With the emergence of facial recognition technology, arguments about whether and how it should be used in routine policing have raised the specter of democracies following China in surveillance of the entire population.[62] Limited use for identification purposes may be more acceptable, but relying on mug shots would replicate the problem with fingerprints and DNA. To that end, a controversial proposal is that the police should have access to no one's biometric data — or everyone's.[63]

B. Unlearning Bias

An alternative approach to the problem of bias in algorithms is to 'unbias' them with regard to specific factors. This draws on one of the advantages algorithms offer over humans: that their decision-making processes can be the subject of experimentation. Whereas an employer who chose to hire a man over a woman is unlikely to admit to bias affecting that specific decision — indeed, there may have been no conscious bias at all — it is possible for algorithms to be run with tweaked parameters to examine whether disparate outcomes would have been reached in different scenarios.[64] That can only be

[59] Gandy, *supra* note 22, at 33-34.

[60] An alternative approach is to seek not to predict future behavior but to shape it. Causal inference is one such approach, in which the goal would be not to categorize offenders such as Eric Loomis into risk groups but to minimize the risk of reoffending through individualized assessment and experimentation: Barabas et al., *supra* note 41, at 6-8. *See generally* GUIDO W. IMBENS AND DONALD B. RUBIN, CAUSAL INFERENCE FOR STATISTICS, SOCIAL, AND BIOMEDICAL SCIENCES: AN INTRODUCTION (2015).

[61] SIMON CHESTERMAN, ONE NATION UNDER SURVEILLANCE: A NEW SOCIAL CONTRACT TO DEFEND FREEDOM WITHOUT SACRIFICING LIBERTY 257-58 (2011).

[62] Daithí Mac Síthigh and Mathias Siems, *The Chinese Social Credit System: A Model for Other Countries?*, _ MOD. L. REV. (forthcoming).

[63] Barry Friedman and Andrew Guthrie Ferguson, *Here's a Way Forward on Facial Recognition*, N.Y. TIMES, Oct. 31, 2019.

[64] Kleinberg et al., *supra* note 22. *See also* Amit Datta, Michael Carl Tschantz, and Anupam Datta, *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination*, 2015(1) PROCEEDINGS ON PRIVACY ENHANCING TECHNOLOGIES 92 (2015). *Cf.* Bambauer and Zarsky, *supra* note 21 (discussing gamification as a strategy for dealing with algorithms); Cofone, *supra* note 22 (noting the paradox that to avoid disparate

done, however, if they are made available to auditors or external testers.

Even then, it may be challenging to define what factors would amount to impermissible bias and how to test for it. Obvious candidates would be those protected by national anti-discrimination laws, such as sex/gender, race, age, religion, disability, and so on.[65] Searching for bias may pose difficulties if there is no baseline against which to measure. Machine learning processes often split data prior to use into training data and validation data. Though that might seem to offer an opportunity to check for bias, the data used to test performance of the model may have the same bias as that used to train it.[66] Even good faith efforts to use algorithms to combat bias may fail if they are unable to take account of social context. 'Fairness', for example, is not a property of a technical system, but of the society within which that system functions.[67]

One of the grounds raised by Eric Loomis in his appeal against the sentencing decision was that COMPAS took gender into account in considering an offender's risk of recidivism. He conceded that men might generally have higher recidivism and violent crime rates than women, but argued that it was a violation of his due process rights to apply that statistical evidence to his case in particular. The court cited some of the literature on the topic and concluded that the use of gender by COMPAS 'promotes accuracy that ultimately inures to the benefit of the justice system including defendants'; in any event, it held, Loomis had not shown that gender was actually relied on as a factor in his sentencing.[68] Discharging that burden was not helped by the fact that, as the court had earlier observed, the algorithm's proprietary nature meant that there was some uncertainty as to *how* gender had been taken into account at all.[69]

---

treatment, protected categories cannot be considered; but to avoid disparate *impact*, they must be).

[65] *See generally* TARUNABH KHAITAN, A THEORY OF DISCRIMINATION LAW (2015).

[66] Karen Hao, *This Is How A.I. Bias Really Happens — and Why It's so Hard to Fix*, MIT TECH. REV., Feb. 4, 2019.

[67] Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, 1(1) ACM CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FAT*) (2018); RICHARD BERK, MACHINE LEARNING RISK ASSESSMENTS IN CRIMINAL JUSTICE SETTINGS 115-30 (2019). *Cf.* Ajunwa, *supra* note 22.

[68] *State v. Loomis*, 767.

[69] *Id*. at 765. *See further* Melissa Hamilton, *The Sexist Algorithm*, 37 BEHAV. SCI. & L. 145 (2019) (arguing that COMPAS systemically over-classifies women in higher risk groupings).

III. ILLEGITIMATE DECISIONS

Opacity, then, may allow inferior decisions or mask impermissible ones. These are matters to mitigate or correct. In a third class of decision-making processes, however, opacity is problematic because the transparency of that process itself may be as important as the effectiveness or appropriateness of the outcome.[70]

Reasoned decision-making on the part of public actors is often said to be foundational to modern notions of liberalism.[71] Much of the literature critiquing algorithmic decision-making by such actors tends to focus on the quality of such decisions, including the possibility of poor decisions due to incomplete or corrupted data,[72] lack of capacity to supervise the relevant systems,[73] or regulatory capture by industry.[74] Alternatively, criticism highlights the discriminatory impact or impermissible bias of such decisions.[75]

These largely reproduce issues discussed in the prior sections of this article. Here, the focus is on two classes of decisions in which opacity itself — as distinct from what it may obscure — undermines legitimacy. The first is in decisions by public actors whose authority is tied to democratic processes that would be frustrated by opacity. The second is in decisions by courts, whose claim to the rule of law depends on public justifications that are intelligible to the wider community: justice being done, but also *seen* to be done.

A. Public Decisions

Edward Shils, a U.S. sociologist writing in the 1950s not long after the McCarthy hearings, argued that liberal democracy depended on protecting privacy for individuals and denying it to government.[76] Succeeding decades have seen the opposite happen: individual privacy has evaporated while

---

[70] *Cf.* Brennan-Marquez, *supra* note 23 (arguing that the threshold of probable cause required by the U.S. Fourth Amendment requires police to account for their decisions, rather than to rely on statistics).

[71] *See, e.g.,* JOHN RAWLS, POLITICAL LIBERALISM (1996); Jeremy Waldron, *Theoretical Foundations of Liberalism*, 37(147) THE PHILOSOPHICAL QUARTERLY 127 (1987).

[72] *See, e.g.,* Barocas and Selbst, *supra* note 22, at 689.

[73] *See, e.g.,* Valentine, *supra* note 22, at 372-75.

[74] *Cf.* John Finch, Susi Geiger, and Emma Reid, *Captured by Technology? How Material Agency Sustains Interaction Between Regulators and Industry Actors*, 46(1) RESEARCH POLICY 160 (2017).

[75] Hacker and Petkova, *supra* note 23, at 7-9. *See also* sourced cited *supra* note 22.

[76] EDWARD A. SHILS, THE TORMENT OF SECRECY: THE BACKGROUND AND CONSEQUENCES OF AMERICAN SECURITY POLICIES 21-25 (1956).

governments have become ever more secretive.[77] Opacity in decision-making is not the same as secrecy, yet it has an analogous effect in undermining the possibility of being held to account for those decisions. It may, arguably, be worse than secrecy because some part of government at least has access to details of classified activities, even if they are not released to the public. Indeed, it is telling that in several cases public bodies have kept the use of opaque algorithms itself a secret.[78]

This form of opacity applies at the micro- as well as the macro-level. At the micro-level, the development of algorithms involves a great many decisions that are political as well as technical. Fine-tuning of parameters may include determinations that privilege one set of interests over another, or affect how public resources are allocated.[79] Accounting for false negatives and positives determines who bears the risk of error, with many instances showing that governments effectively transferred that risk to their most vulnerable citizens in areas ranging from welfare benefits to probation determinations and foster care.[80]

In the United States, a handful of lawsuits have been successful in challenging opaque government decisions relating to discontinuation of benefits and the sacking of public school teachers, relying on due process protections under the Fourteenth Amendment.[81] Greater protections are found in the European Union, though these are typically linked to safeguards against being subject to *automatic* processing, rather than being the subject of *opaque* decision-making as such.[82]

The EU's 1995 Data Protection Directive gave individuals rights to obtain information about whether and how their personal data was processed, including the right to obtain 'knowledge of the logic involved in any automatic processing'.[83] That provision applied to public and private sector decisions, but does not seem to have been the subject of significant debate or

---

[77] CHESTERMAN, *supra* note 61, at 67-89.

[78] Valentine, *supra* note 22, at 376-78.

[79] Mittelstadt et al., *supra* note 39.

[80] *See, e.g.,* Jason Parkin, *Adaptable Due Process*, 160 U. PA. L. REV. 1309, 1357-58 (2012) (welfare benefits); Brauneis and Goodman, *supra* note 23, at 120 (probation decisions); VIRGINIA EUBANKS, AUTOMATING INEQUALITY: HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR 144-55 (2017) (foster care).

[81] Valentine, *supra* note 22, at 413-19.

[82] European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment (European Commission for the Efficiency of Justice (CEPEJ), Strasbourg, Dec. 4, 2018), *at* http://www.coe.int/cepej. *Cf.* the separate 'right to good administration' recognized under EU law: DAMIAN CHALMERS, GARETH DAVIES, AND GIORGIO MONTI, EUROPEAN UNION LAW: TEXT AND MATERIALS 377-79 (4th ed. 2019).

[83] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data (EU Data Protection Directive) 1995 (EU), art 12(a).

litigation.[84] With the adoption of the GDPR in 2016, it was expanded to include a right of access to 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing'.[85]

The new language coincided with growing awareness of the opacity of many algorithmic processes. Whether it amounts to a 'right to explanation' has been the subject of much discussion.[86] Of particular interest is the import of the word 'meaningful'.[87] The EU Working Party on the topic appears to have aligned itself with the more limited interpretation, observing that the provision requires that subjects be provided with 'information about the *envisaged consequences* of the processing, rather than an explanation of a *particular* decision'. [88] Acknowledging the difficulties imposed by complexity, those providing the information are enjoined to find 'simple ways to tell the data subject about the rationale behind, or the criteria relied on in reaching the decision' — which need not include a 'complex explanation of the algorithm used' or disclosure of the algorithm itself.[89]

A further constraint is that the right to explanation (if it exists) is limited by its connection to the right not to be subject to automated processing. That is, the GDPR limits autonomous decision-making processes — including those that are opaque — but does not apply directly to decision-making processes in which a human is supported by algorithms that may themselves be opaque. [90] The GDPR also allows automated processing where it is necessary for a contract, authorized by law, or based on the subject's 'explicit consent'.[91] Final restrictions to these rights come in the form of carve-outs. A recital states that the right of access should not adversely affect 'the rights

---

[84] Lilian Edwards and Michael Veale, *Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?*, 16(3) IEEE SECURITY & PRIVACY 46, 47 (2018).

[85] GDPR, *supra* note 35, art 15(1)(h).

[86] Bryce Goodman and Seth Flaxman, *European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"*, 38(3) A.I. MAG. 50 (2017); Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, *Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation*, 7(2) INT'L DATA PRIVACY L. 76 (2017); Andrew D. Selbst and Julia Powles, *Meaningful Information and the Right to Explanation*, 7 INT'L DATA PRIVACY L. 233 (2017); Casey, Farhangi, and Vogl, *supra* note 23.

[87] Michael Veale and Lilian Edwards, *Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making And Profiling*, 34(2) COMPUTER L. & SEC. REV. 398, 399-400 (2018).

[88] Guidelines on Automated Individual Decision-Making, *supra* note 36, at 27 (emphasis in original).

[89] *Id*. at 25.

[90] Edwards and Veale, *supra* note 84, at 47. *See supra* note 36.

[91] GDPR, *supra* note 35, art 22(2).

or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software.'[92] And, though the GDPR applies to both public as well as private sector decisions, it expressly excludes data processing by competent authorities for the purposes of preventing, investigating, and prosecuting criminal offences.[93]

More expansive protection is offered in a 2016 French law, which created a right to request information about algorithmic decisions made by administrative bodies, including the rules and main characteristics of the algorithm.[94] A subsequent decree elaborated that the information was to include the parameters of the algorithm as well as their weighting, and that it should be in 'intelligible form'.[95] This last provision points to one of the key limitations of 'explanation' or 'transparency' as the remedy to opacity. Providing information in a manner that is intelligible to the average person, yet complete enough to give a full explanation of an algorithmic process, while not unreasonably compromising trade secrets or allowing users to game the system, is exceedingly difficult.[96]

A more effective remedy may, in fact, be traditional administrative law. If, for example, a decision-maker is not permitted to delegate a decision to a third party, he or she should not be able to delegate it to an A.I. system; if the decision-maker is given discretion, that discretion should not be unlawfully fettered.[97] Though there is no general duty to give reasons for all decisions, such a duty is often imposed by statute, or by the common law where the decision is judicial or quasi-judicial in nature.[98] If the use of an A.I. system precluded the giving of such reasons, judicial review might conclude that the decision was irrational, or impugnable on the basis that it could not be shown

---

[92] *Id*., recital 63. *Cf.* Selbst and Powles, *supra* note 86, at 242 (arguing that this provision should be read down in light of other changes in the GDPR).

[93] GDPR, *supra* note 35, art 2(2)(d).

[94] Loi no 2016-1321 du 7 octobre 2016 pour une République numérique 2016 (France), art 4: 'une décision individuelle prise sur le fondement d'un traitement algorithmique comporte une mention explicite en informant l'intéressé. Les règles définissant ce traitement ainsi que les principales caractéristiques de sa mise en œuvre sont communiquées par l'administration à l'intéressé s'il en fait la demande.' *See generally* Constance Chevallier-Govers, *Right of Access to Public Documents in France*, *in* THE RIGHT OF ACCESS TO PUBLIC INFORMATION: AN INTERNATIONAL COMPARATIVE LEGAL SURVEY 265, 271 (Hermann-Josef Blanke and Ricardo Perlingeiro eds., 2018).

[95] Décret n° 2017-330 du 14 mars 2017 relatif aux droits des personnes faisant l'objet de décisions individuelles prises sur le fondement d'un traitement algorithmique 2017 (France), art 1. *See also* Edwards and Veale, *supra* note 84, at 48-49.

[96] Wachter, Mittelstadt, and Russell, *supra* note 21, at 842-43.

[97] Jennifer Cobbe, *Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making*, 39 LEGAL STUDIES 636, 644-47 (2019).

[98] *Id*. at 648.

whether material factors were taken into account and that immaterial factors were not.[99]

A residual problem, however, is the Catch-22 of opacity: efforts to challenge such decisions are hampered by the very opacity that might form the basis of an action — people do not know what they don't know. In any case, relying upon individuals to request transparency means that it will only be the most motivated who do so.[100] The hypothetical right to explanation may, then, end up serving the same function as consent in data protection law: a formal basis for legitimacy in theory, though untethered from any meaningful agreement between equals in practice.[101]

B. Courts

Attempts to restrain opaque decision-making by public bodies will be limited in their effectiveness, in part because the default posture of many such entities is to give reasons only when asked. Not so courts and related tribunals, where reasons are expected as a matter of course.

That is not to say that courts never rely on metaphorical black boxes themselves. Juries are the most prominent example. In those jurisdictions where they are used, jurors reach verdicts in civil and criminal cases without providing reasons. They are meant to be guided by the judge, however, who often retains the power to ignore their verdict if he or she determines that no 'reasonable' jury could have reached it.[102]

As a growing portion of the criminal justice system comes to rely upon technology, these problems are going to increase. From predictive policing models to forensic software programs used in trials, algorithms protected as trade secrets are now used at all stages of criminal proceedings.[103] One

---

[99] *Id*. at 650-51. *Cf.* Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008).

[100] Edwards and Veale, *supra* note 84, at 49.

[101] Simon Chesterman, *Introduction*, *in* DATA PROTECTION LAW IN SINGAPORE: PRIVACY AND SOVEREIGNTY IN AN INTERCONNECTED WORLD, 2-3 (Simon Chesterman ed., 2018). In theory, when personal data is collected, used and disclosed, that is in accordance with the agreement of the identifiable individual concerned, often through a contractual arrangement with an organization. In practice, of course, users confronted with multi-page end-user license agreements simply click "I accept" and get on with whatever they wanted to do in the first place. The British retailer GameStation provided a memorable example of this one April Fool's Day, when more than 7,000 people clicked "I accept" to terms and conditions that included the surrender of their immortal souls to the company. (The company later rescinded all claims, temporal and spiritual.)

[102] *Cf.* Jason Iuliano, *Jury Voting Paradoxes*, 113 MICH. L. REV. 405 (2014).

[103] *See, e.g.,* Rashida Richardson, Jason M. Schultz, and Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 N.Y.U. L. REV. 192 (2019).

response would be to abolish the trade secrets privilege in criminal trials, essentially forcing companies to reveal how such technologies reach their conclusions. [104] Alternatively, some courts have excluded evidence completely where opacity renders its use suspect. [105] It is unclear how effective such safeguards will be, given the internal and external pressures on judges to use such assessments and their relative inexperience in evaluating such tools. [106]

A vision of the future in western courts may be offered by the extensive use of technology in the Chinese legal system. China's automated surveillance of its population, including the 'social credit system', has been much reported. [107] Less recognized is the manner in which algorithms now support the Chinese legal system. The Judicial Accountability System [司法责任制] began as a campaign to promote consistency in judgments. Past efforts had relied on reviews by superiors, but this was impractical and undermined the authority of the judge who heard the case. [108] In its place, judges are now required to search for similar cases prior to making a judgment. This is said to have led to experiments in local courts with A.I. systems that 'push' similar cases up to a judge prior to him or her taking a decision, or flag an 'abnormal judgment warning' if the proposed judgment departs significantly from other cases. [109] These are part of a suite of technologies that have been adopted, influenced both by the supply of technology companies in China and the demands of a complex and developing legal system. [110] In 2017, the Wujiang District of Suzhou trialed a 'one-click' summary judgment process, which automatically generated proposed grounds of decision complete with sentence. [111] This now appears

[104] Wexler, *supra* note 21 (arguing that trade secrets should not be privileged in criminal proceedings as this overprotects intellectual property rights at the expense of due process rights of the accused).

[105] *People v Fortin*, 218 Cal Rptr 3d 867 (California Court of Appeals, 2017) (excluding use of software introduced by the defendant in a sexual molestation case, inter alia because the developer refused to share the underlying formula with other scientists, undermining its evidentiary value).

[106] Note, *State v. Loomis: Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing*, 130 HARV. L. REV. 1530 (2017).

[107] *See* Mac Síthigh and Siems, *supra* note 62.

[108] *Cf.* Margaret Y.K. Woo, *Court Reform with Chinese Characteristics*, 27 WASH. INT'L L.J. 241 (2017) (describing the manner in which workload and the impact of reversals on a judge's career encouraged dismissal of cases on technical grounds).

[109] YU Meng and DU Guodong, *Why Are Chinese Courts Turning to A.I.?*, THE DIPLOMAT, Jan. 19, 2019.

[110] Masha Borak, *China Embraces Tech in Its Courtrooms*, TECH NODE, Oct. 24, 2018, *at* http://technode.com/2018/10/24/china-court-technology/.

[111] *苏州法院刑案简易判决一键生成 [One-click Generation of the Summary Judgment of the Criminal Case in Suzhou Court]*, 法制日报 [LEGAL DAILY], June 19, 2017.

to be expanding to the wider court system.[112]

In a speech in early 2019, Singapore's Chief Justice noted that such developments in China and elsewhere were making 'machine-assisted court adjudication a reality'. At the same time, he noted, the use of A.I. within a justice system gives rise to a 'unique set of ethical concerns, including those relating to credibility, transparency and accountability'.[113] To this one might add considerations of equity, since the drive the greater automation in civil proceedings is being dominated by deep-pocketed clients with uncertain consequences for the future administration of justice.[114]

Uncertainty about the appropriate checks and balances to manage those concerns has led to some knee-jerk responses. In 2019, for example, France — again an outlier in saying a loud '*non*' to algorithms — adopted an extraordinary law prohibiting the publication of data analytics that reveal or predict how particular judges decide on cases. Punishable by jail time, the new offence was reportedly adopted after considering an alternative that would have seen judgments published without identifying judges by name at all.[115]

Elsewhere, judges continue to muddle along. In practice, the barriers to a successful challenge to the use of algorithms in a courtroom are likely to be high, as Eric Loomis found out. His appeal against the circuit court's sentencing decision on the basis that his due process rights had been violated was unsuccessful. The Wisconsin Supreme Court conceded that defendants are entitled to be sentenced based on accurate information, but it was enough that he had the opportunity to verify the answers he gave when COMPAS calculated its score. As for the score itself, it was not true that the circuit court had relied on information to which Loomis was denied access — for Judge Horne himself also had no knowledge of how the score had been reached.[116]

The Wisconsin Supreme Court ultimately upheld the decision, finding that consideration of the COMPAS score was supported by other independent factors and 'not determinative' of his sentence.[117] It went on, however, to

---

[112] DU Guodong and YU Meng, *How China's E-Justice System Works?*, CHINA LAW & PRACTICE, Jan. 19, 2019.

[113] Sundaresh Menon, *Opening of the Legal Year* (Singapore, Supreme Court, Jan. 7, 2019), *at* http://www.supremecourt.gov.sg/docs/default-source/default-document-library/cj-oly-speech-2019-pdf.pdf at 10-11.

[114] Endo, *supra* note 21. On the transformation of the legal profession more generally, see *supra* note 20.

[115] Loi no 2019-222 du 23 mars 2019 de programmation 2018-2022 et de réforme pour la justice 2019 (France), art 33; *France Bans Judge Analytics, 5 Years In Prison For Rule Breakers*, ARTIFICIAL LAWYER, June 4, 2019, *at* http://www.artificiallawyer.com/2019/06/04/france-bans-judge-analytics-5-years-in-prison-for-rule-breakers.

[116] *State v. Loomis*, 760-61

[117] *Id*. at 753.

express reservations about the use of such software, requiring that future use must be accompanied by a 'written advisement' about the proprietary nature of the software and the limitations of its accuracy.[118] Chief Justice Roggensack added a concurrence in which she clarified that a court may *consider* tools like COMPAS in sentencing but must not *rely* on them.[119] A fellow justice went further, arguing that sentencing decisions should include a record explaining the limitations of such systems as part of the 'long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing'.[120] The U.S. Supreme Court declined to hear an appeal.[121]

CONCLUSION

'Publicity,' Jeremy Bentham wrote more than two centuries ago, 'is the very soul of justice. … It keeps the judge himself, while trying, under trial.'[122] Judicial decisions are the clearest example of an area in which the use of opaque A.I. systems should be limited, but even there we see 'algorithm creep'. As this article has shown, computational methods have introduced efficiencies and optimization to a wide range of decision-making processes — though at a cost.

In some cases, the trade-off is worthwhile. Where output-based legitimacy is sufficient, ignorance may not be bliss, but it is tolerable. The choice to use an opaque system itself, however, should be a conscious and informed one. That choice should include consideration of the risks that come

---

[118] *Id*. at 763-64: 'Specifically, any PSI containing a COMPAS risk assessment must inform the sentencing court about the following cautions regarding a COMPAS risk assessment's accuracy: (1) the proprietary nature of COMPAS has been invoked to prevent disclosure of information relating to how factors are weighed or how risk scores are to be determined; (2) risk assessment compares defendants to a national sample, but no cross-validation study for a Wisconsin population has yet been completed; (3) some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism; and (4) risk assessment tools must be constantly monitored and re-normed for accuracy due to changing populations and subpopulations. Providing information to sentencing courts on the limitations and cautions attendant with the use of COMPAS risk assessments will enable courts to better assess the accuracy of the assessment and the appropriate weight to be given to the risk score.'

[119] *Id*. at 774.

[120] *Id*. at 775.

[121] Cert. denied, 582 U.S. _ (June 26, 2017) (No. 16-6387).

[122] Jeremy Bentham, *Draught for the Organization of Judicial Establishments (1790)*, *in* IV THE WORKS OF JEREMY BENTHAM 285, 316 (John Bowring ed., 1843).

with opacity.

The regulatory response to this opacity has been inconsistent. That is often the case with new technologies. Writing in 1980, David Collingridge observed that any efforts at control face a double bind. During the early stages, when control would be possible, not enough is known about the technology's harmful social consequences to warrant slowing its development; by the time those consequences are apparent, control has become costly and slow.[123] European efforts to restrain automatic processing clearly weigh the harmful social consequences more heavily than they are perceived in China. The United States experience of predictive sentencing, for its part, exemplifies the difficulty of reining in a technology whose use has effectively become standard.

It is often presumed that the remedy to opacity is transparency. Yet this article has argued that the problem of opacity should be understood in three discrete ways: such decisions may be inferior, they may mask impermissible biases, or they may be illegitimate merely because of their opacity. Each points to slightly different remedies.

Poor decisions may be improved by more robust testing and verification. Success might be measured in the quality of those decisions, a cost-benefit analysis viewed through a utilitarian lens. Avoiding bias, by contrast, may benefit from greater clarity as to how and why algorithms are used. The goal should not be mere optimization, but appropriate weighing of social and cultural norms, with rigorous audits to ensure that these are not being compromised.[124] Success here is more complicated, as discrimination law rarely offers bright lines comparable to, say, the proposed ban on allowing algorithms to control lethal weapons.[125]

In a third class of cases, the need to explain a decision is a kind of process legitimacy, applicable especially where public authorities take decisions affecting the rights and obligations of individuals. The inability to explain how such a decision was made will, in some circumstances, be akin to the decision itself having been impermissibly delegated to another party. Success here most closely tracks the calls for transparency and explainability in A.I. systems — though primarily so that a human decision-maker can still be held accountable for those decisions.

In the course of Eric Loomis's appeal, the Wisconsin assistant attorney general representing the state implicitly questioned whether that was, in fact, such an important shibboleth. After all, she said, 'We don't know what's

---

[123] DAVID COLLINGRIDGE, THE SOCIAL CONTROL OF TECHNOLOGY 19 (1980).

[124] Some have gone further to suggest that these could be used for progressive purposes, a kind of 'algorithmic affirmative action'. Chander, *supra* note 22, at 1039-45.

[125] *See* Chesterman, *supra* note 11.

going on in a judge's head; it's a black box, too.'[126] As for Mr Loomis himself, he was released from Jackson Correctional Institution in August 2019 after serving his full six-year term.[127] According to COMPAS, at least, there is a high risk he will return.

[126] Jason Tashea, *Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions*, A.B.A. J., Mar. 1, 2017 (quoting Christine Remington).

[127] State of Wisconsin Department of Corrections Offender Locator, *at* https://appsdoc.wi.gov/lop.