3

Opacity

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 2am on a Monday morning in February 2013. Though no one was injured, police were called and soon identified Loomis's Dodge Neon two kilometres away. A short car chase ended when he crashed into a snowdrift; together with a passenger he continued on foot, but he 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.

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.[1] 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,[2] 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

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

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

https://doi.org/10.1017/9781009047081.005 Published online by Cambridge University Press

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'.[3]

Opacity is the antithesis of legal decisions. Accountability for those decisions typically requires that the decision-maker have a convincing reason for a decision or act. Judicial decisions in particular give special weight to reasoning.[4] 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.[5] 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 US Supreme Court.

The problem of understanding AI systems is not new. 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.[6] More prosaically, it has been argued that computer simulation displaces humans from the centre 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'.[7] Distinct from the challenges posed by autonomy in AI systems, the increasing opacity of those systems is not a challenge to the centrality of human agents as legal *actors* so much as to our ability to

[3] *State v Loomis*, 881 NW 2d 749, 755 (Wis, 2016).
[4] Herbert Wechsler, 'Toward Neutral Principles of Constitutional Law' (1959) 73 Harvard Law Review 1, 19–20.
[5] 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' (2015) 72 Washington & Lee Law Review 483.
[6] Frank Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015) 190.
[7] Paul Humphreys, 'The Philosophical Novelty of Computer Simulation Methods' (2009) 169 Synthese 615, 617.

understand and evaluate *actions* – something essential to meaningful regulation.

'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 arises in complex systems that require specialist skills to understand them. These systems evolve over time, being added to by different stakeholders, but are in principle capable of being explained.

Neither of these forms of opacity – proprietary or complex – poses new problems for law. Intellectual property law has long recognized protection of intangible creations of the human mind and exceptions based on fair use.[8] To deal with complex issues, governments and judges routinely have recourse to experts.[9] 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.[10]

To pick an example mentioned in the introduction to this book, 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.[11]

[8] Amanda Levendowski, 'How Copyright Law Can Fix Artificial Intelligence's Implicit Bias Problem' (2018) 93 Washington Law Review 579. On the question of intellectual property created by an AI system itself, see chapter five, section 5.2.2.

[9] See, eg, Carol AG Jones, *Expert Witnesses: Science, Medicine, and the Practice of Law* (Clarendon Press 1994).

[10] For a description of machine learning, see the introduction to this book at n 2. Cf Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society. '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.

[11] 'Google's AI Beats World Go Champion in First of Five Matches', *BBC News* (9 March 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' (2017) 550 Nature 354.

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 a barrier to allowing it into the market.[12] 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.[13] 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.

There is a growing literature criticizing reliance on algorithmic decision-making with legal consequences. A significant portion now focuses on opacity, highlighting specific concerns such as bias, or seeking remedies through transparency. 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 EU, as well as less-studied innovations in China, this chapter develops a typology of those challenges 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 behaviour; they also encourage good behaviour. 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

[12] Alex John London and Jonathan Kimmelman, 'Why Clinical Translation Cannot Succeed without Failure' (2015) 4 eLife e12844. 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* (Norton 2019).

[13] 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' (2019) 25 Nature Medicine 16.

prison sentences, for example, acting on such predictions would not be tolerated in a judge and should not be accepted in an AI 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.

It will be apparent that this structure echoes the discrete challenges raised about autonomy in chapter two: the quality of outcomes approaches the question through a utilitarian lens; the avoidance of impermissible decisions reflects deontic concerns; while reliance upon proper authority and process is sought to confer legitimacy – in this case based not on the identity of the actor but on the publicness of his or her reasoning.

The means of addressing some or all of these concerns is routinely said to be through transparency.[14] 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.

3.1 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.[15] Uber similarly designed a version of its app that identified users whose behaviour suggested that they were working for regulators in order to limit their ability to gather evidence.[16]

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.

[14] See chapter six.

[15] EPA, California Notify Volkswagen of Clean Air Act Violations/Carmaker Allegedly Used Software That Circumvents Emissions Testing for Certain Air Pollutants (US Environmental Protection Agency, 18 September 2015).

[16] Leslie Hook, 'Uber Used Fake App to Confuse Regulators and Rivals', *Financial Times* (4 March 2017).

https://doi.org/10.1017/9781009047081.005 Published online by Cambridge University Press

In others, the patchiness of data might be revealed, raising questions about the reliability of the process or the confidence level of the outcome.[17] This phenomenon of 'garbage in, garbage out' is as old as the first computer. Charles Babbage, the English polymath who fashioned the 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.[18]

Human complacency and automation bias make these more than theoretical problems. As human involvement in a process – notionally 'in' or 'over' the loop[19] – is reduced to its most mechanistic, the tendency to accept default suggestions increases.[20] This may be compared with the danger, discussed in chapter two, 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.[21] 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.[22]

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 behaviour, they were also favoured because using them minimized the risk that the lawyers and judges themselves would be blamed for the consequences of their decisions.[23]

[17] Sandra Wachter and Brent Mittelstadt, 'A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI' [2019] Columbia Business Law Review 494.

[18] Charles Babbage, *Passages from the Life of a Philosopher* (Longman 1864) 67.

[19] See chapter two, section 2.3.

[20] Steven PR Rose and Hilary Rose, '"Do Not Adjust Your Mind, There Is a Fault in Reality" – Ideology in Neurobiology' (1973) 2 Cognition 479, 498–99. On the larger impact of anchoring in sentencing decisions, see Birte Enough and Thomas Mussweiler, 'Sentencing under Uncertainty: Anchoring Effects in the Courtroom' (2001) 31 Journal of Applied Social Psychology 1535.

[21] See chapter two, section 2.1.

[22] Raja Parasuraman and Dietrich Manzey, 'Complacency and Bias in Human Use of Automation: An Attentional Integration' (2010) 52 Human Factors 381, 392.

[23] Kelly Hannah-Moffat, 'The Uncertainties of Risk Assessment: Partiality, Transparency, and Just Decisions' (2015) 27 Federal Sentencing Reporter 244.

Addressing complacency and automation bias goes far beyond the regulatory challenges that are the focus of this book. 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 EU offers comparatively stronger protections under its GDPR, which makes clear that the right not to be subject to automated processing[24] 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'.[25]

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.[26] 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.

3.2 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.[27] They are only as good as the data they are given and the questions they are asked, however. In

[24] See chapter two, section 2.3.2.

[25] 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, 3 October 2017) 20–21.

[26] Sheen S Levine and Michael J Prietula, 'Open Collaboration for Innovation: Principles and Performance' (2014) 25 Organization Science 1287.

[27] Sharad Goel et al, 'Combatting Police Discrimination in the Age of Big Data' (2017) 20 New Criminal Law Review 181. This may be particularly useful in decision-making systems that are delegated and distributed: Katherine J Strandburg, 'Rulemaking and Inscrutable Automated Decision Tools' (2019) 119 Columbia Law Review 1851, 1857.

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.

A prominent example is screening decisions. Many industries now use AI 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.

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, like fairness, or when algorithms are used in order to predict future behaviour by specific individuals,[28] 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 two most important factors indicative of job performance at a particular company were being named Jared and having played high-school lacrosse.[29] In others, reliance upon algorithms may reflect or reify discriminatory practices.

3.2.1 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. 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

[28] Chelsea Barabas et al, 'Interventions over Predictions: Reframing the Ethical Debate for Actuarial Risk Assessment' (2018) 81 Proceedings of Machine Learning Research 1.

[29] Dave Gershgorn, 'Robot Indemnity: Companies Are on the Hook if Their Hiring Algorithms Are Biased', *Quartz* (22 October 2018).

which outputs are interpreted.[30] Various scholars compare this to the distinction between 'disparate treatment', or intentional behaviour, and 'disparate impact' in US civil rights jurisprudence.[31] 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 treated less favourably than men's.[32]

Examples of unintended bias include facial recognition software that is less effective at recognizing dark-skinned faces because its training was done using light-skinned ones.[33] The use of unrepresentative data is not unique to AI 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.[34] 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 maternity leave.[35]

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 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'.[36] The report was criticized for oversimplifying

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

[31] *Ricci v DeStefano*, 557 US 557 (2009). See, eg, Solon Barocas and Andrew D Selbst, 'Big Data's Disparate Impact' (2016) 104 California Law Review 671, 694–712.

[32] Ignacio N Cofone, 'Algorithmic Discrimination Is an Information Problem' (2019) 70 Hastings Law Journal 1389, 1397–98.

[33] Karl Manheim and Lyric Kaplan, 'Artificial Intelligence: Risks to Privacy and Democracy' (2019) 21 Yale Journal of Law & Technology 106, 159.

[34] Joseph Henrich, Steven J Heine, and Ara Norenzayan, 'The Weirdest People in the World?' (2010) 33 Behavioral and Brain Sciences 61 (the title refers to subjects being drawn entirely from Western, Educated, Industrialized, Rich, and Democratic (WEIRD) societies).

[35] Cf Rafael Lalive et al, 'Parental Leave and Mothers' Careers: The Relative Importance of Job Protection and Cash Benefits' (2014) 81 Review of Economic Studies 219.

[36] Julia Angwin et al, 'Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks', *ProPublica* (23 May 2016).

risk assessment, cherry-picking results, and ignoring the higher incarceration rates of African Americans.[37] 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.[38]

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.[39] 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 is not cause.

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 enable the identification of optimal weights for those variables that best predict the outcome of interest.[40] 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.[41] The algorithm's impenetrability, however, and the criticism to which that gave rise anticipate future challenges as AI systems become more complex and play a greater role in decisions affecting the rights and obligations of individuals.

[37] 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"' (2016) 80 (2) Federal Probation 38.

[38] Alexandra Chouldechova, 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments' (2017) 5 Big Data 153.

[39] Malcolm M Feeley and Jonathan Simon, 'The New Penology: Notes on the Emerging Strategy of Corrections and Its Implications' (1992) 30 Criminology 449; Paula Maurutto and Kelly Hannah-Moffat, 'Assembling Risk and the Restructuring of Penal Control' (2006) 46 British Journal of Criminology 438.

[40] Andrew Gelman and Jennifer Hill, *Data Analysis Using Regression and Multilevel/Hierarchical Models* (Cambridge University Press 2007).

[41] A Practitioner's Guide to COMPAS Core (Northpointe, 2015) 29.

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.[42] Here opacity becomes a concern as the black box nature 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.[43] 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 behaviour. 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 are considered when choosing from a range of possible sentences, or when considering early release, truly preventive detention is rare in most well-ordered jurisdictions.[44] The second objection is that the application of summary statistics to individuals is the very definition of stereotyping.[45] 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.[46]

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

[43] See, eg, Kelly Hannah-Moffat, 'Sacrosanct or Flawed: Risk, Accountability and Gender-Responsive Penal Politics' (2011) 22 Current Issues in Criminal Justice 193; Seth J Prins and Adam Reich, 'Can We Avoid Reductionism in Risk Reduction?' (2018) 22 Theoretical Criminology 258. Mental health, for example, tends to be excluded in favour of more measurable and statistically significant covariates.

[44] Hallie Ludsin, *Preventive Detention and the Democratic State* (Cambridge University Press 2016).

[45] Oscar H Gandy, Jr, 'Engaging Rational Discrimination: Exploring Reasons for Placing Regulatory Constraints on Decision Support Systems' (2010) 12 Ethics and Information Technology 29, 33–34.

[46] An alternative approach is to seek not to predict future behaviour 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. See generally Guido W Imbens and Donald B Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press 2015).

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.[47] With the emergence of facial recognition technology, arguments about whether and how it should be used in routine policing have raised the spectre of democracies following China in surveillance of the entire population. 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.[48]

3.2.2 *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 that algorithms offer over humans: 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.[49] That can only be done, however, if they are made available to auditors or external testers.[50]

One of the grounds raised by Eric Loomis in his appeal against the sentencing decision was that COMPAS took gender into account when 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

[47] Simon Chesterman, *One Nation under Surveillance: A New Social Contract to Defend Freedom without Sacrificing Liberty* (Oxford University Press 2011) 257–58.

[48] Barry Friedman and Andrew Guthrie Ferguson, 'Here's a Way Forward on Facial Recognition', *New York Times* (31 October 2019).

[49] Jon Kleinberg et al, 'Discrimination in the Age of Algorithms' (2018) 10 Journal of Legal Analysis 113. 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.

[50] See chapter six, section 6.2.2(b).

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. 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 whether gender had been taken into account at all.[51]

3.3 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.

Reasoned decision-making on the part of public actors is often said to be foundational to modern notions of liberalism.[52] Much of the literature critiquing algorithmic decision-making concentrates on the quality of decisions, including the possibility of poor decisions due to incomplete or corrupted data, lack of capacity to supervise the relevant systems, or regulatory capture by industry.[53] Alternatively, criticism highlights the discriminatory impact or impermissible bias of those decisions.[54]

Those arguments rehearse issues discussed in the prior sections of this chapter. Here, the focus is on two classes of decision 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

[51] *State v Loomis* (n 3) 765–67.

[52] See, eg, John Rawls, *Political Liberalism* (Columbia University Press 1996); Jeremy Waldron, 'Theoretical Foundations of Liberalism' (1987) 37(147) The Philosophical Quarterly 127.

[53] Cf John Finch, Susi Geiger, and Emma Reid, 'Captured by Technology? How Material Agency Sustains Interaction between Regulators and Industry Actors' (2017) 46 Research Policy 160. See also chapter eight, section 8.1.2.

[54] Philipp Hacker and Bilyana Petkova, 'Reining in the Big Promise of Big Data: Transparency, Inequality, and New Regulatory Frontiers' (2017) 15 Northwestern Journal of Technology and Intellectual Property 1, 7–9.

justifications that are intelligible to the wider community: justice being done but also being *seen* to be done.

3.3.1 Public Decisions

Edward Shils, a US 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.[55] Succeeding decades have seen the opposite happen: individual privacy has evaporated while governments have become ever more secretive. 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.

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.[56] 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.[57]

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.[58] Greater protections are

[55] Edward A Shils, *The Torment of Secrecy: The Background and Consequences of American Security Policies* (Heinemann 1956) 21–25.

[56] Brent Daniel Mittelstadt et al, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) Big Data & Society.

[57] See, eg, Jason Parkin, 'Adaptable Due Process' (2012) 160 University of Pennsylvania Law Review 1309, 1357–58 (welfare benefits); Robert Brauneis and Ellen P Goodman, 'Algorithmic Transparency for the Smart City' (2018) 20 Yale Journal of Law & Technology 103, 120 (probation decisions); Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's 2017) 144–55 (foster care).

[58] Sarah Valentine, 'Impoverished Algorithms: Misguided Governments, Flawed Technologies, and Social Control' (2019) 46 Fordham Urban Law Journal 364, 413–19.

found in the EU, though these are typically linked to safeguards against being subject to *automatic* processing, rather than being the subject of *opaque* decision-making as such.[59]

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'.[60] That provision applied to public and private sector decisions, but it does not seem to have been the subject of significant debate or litigation. 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'.[61]

The new language coincided with growing awareness of the opacity of many algorithmic processes. Whether it amounts to a 'right to explanation' will be considered in chapter six.[62] Of particular interest is the import of the word 'meaningful'.[63] 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'.[64] 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.[65]

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

[59] European Ethical Charter on the Use of Artificial Intelligence in Judicial Systems and Their Environment (European Commission for the Efficiency of Justice (CEPEJ), 4 December 2018). See chapter two, section 2.3. Cf the separate 'right to good administration' recognized under EU law: Damian Chalmers, Gareth Davies, and Giorgio Monti, *European Union Law: Text and Materials* (4th edn, Cambridge University Press 2019) 377–79.

[60] 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).

[61] General Data Protection Regulation 2016/679 (GDPR) 2016 (EU), art 15(1)(h).

[62] See chapter six, section 6.3.1.

[63] Michael Veale and Lilian Edwards, 'Clarity, Surprises, and Further Questions in the Article 29 Working Party Draft Guidance on Automated Decision-Making and Profiling' (2018) 34 Computer Law & Security Review 398, 399–400.

[64] Guidelines on Automated Individual Decision-Making (n 25) 27 (emphasis in original).

[65] Ibid 25.

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.[66] The GDPR also allows automated processing where it is necessary for a contract, authorized by law, or based on the subject's 'explicit consent'.[67] Final restrictions of these rights come in the form of carve-outs. A recital states that the right of access should not adversely affect 'the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software'.[68] 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.[69]

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 AI system; if the decision-maker is given discretion, that discretion should not be unlawfully fettered. 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. If the use of an AI system precluded the giving of reasons, judicial review might conclude that the decision was irrational, or impugnable on the basis that it could not be shown whether material factors were taken into account and that immaterial factors were not.[70]

A residual problem, however, is the Catch-22 of opacity: efforts to challenge 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 be only the most motivated who do so. The hypothetical right to explanation may, then, end up serving the same function as consent in data

[66] Lilian Edwards and Michael Veale, 'Enslaving the Algorithm: From a "Right to an Explanation" to a "Right to Better Decisions"?' (2018) 16(3) IEEE Security & Privacy 46, 47. See above n 25.

[67] GDPR, art 22(2).

[68] Ibid, recital 63.

[69] Ibid, art 2(2)(d).

[70] Jennifer Cobbe, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' (2019) 39 Legal Studies 636, 650–51. Cf Danielle Keats Citron, 'Technological Due Process' (2008) 85 Washington University Law Review 1249.

protection law: a formal basis for legitimacy in theory, though untethered from any meaningful agreement between equals in practice.[71]

3.3.2 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.[72]

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.[73] One response would be to abolish the trade secrets privilege in criminal trials, essentially forcing companies to reveal how conclusions are reached.[74] Alternatively, some courts have excluded evidence completely where opacity renders its use suspect.[75] It is unclear how effective these measures will be, given the internal and external pressures on judges to use assessments and their relative inexperience in evaluating the tools making them.

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.[76] Less recognized is the manner in which

[71] Simon Chesterman, 'Introduction' in Simon Chesterman (ed), *Data Protection Law in Singapore: Privacy and Sovereignty in an Interconnected World* (2nd edn, Academy 2018) at 2–3.

[72] Cf Jason Iuliano, 'Jury Voting Paradoxes' (2014) 113 Michigan Law Review 405.

[73] Rashida Richardson, Jason M Schultz, and Kate Crawford, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' (2019) 94 New York University Law Review 192.

[74] Rebecca Wexler, 'Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System' (2018) 70 Stanford Law Review 1343.

[75] *People v Fortin*, 218 Cal Rptr 3d 867 (California Court of Appeals, 2017).

[76] See Daithí Mac Síthigh and Mathias Siems, 'The Chinese Social Credit System: A Model for Other Countries?' (2019) 82 Modern Law Review 1034.

algorithms now support the Chinese legal system.[77] 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 – passed 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.[78]

Elsewhere, judges continue to muddle along. In practice, the barriers to a successful challenge to the use of algorithms in a courtroom are 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.[79]

The superior court ultimately upheld the decision, finding that consideration of the COMPAS score was supported by other independent factors and 'not determinative' of his sentence. It went on, however, to 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.[80] 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. A fellow justice went further, arguing that sentencing decisions should include a record explaining their limitations as part of the 'long-standing, basic requirement that a circuit court explain its exercise of discretion at sentencing'.[81] The US Supreme Court declined to hear an appeal.[82]

[77] See chapter nine, introduction.
[78] 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* (4 June 2019).
[79] *State v Loomis* (n 3) 760–61.
[80] Ibid 753–64.
[81] Ibid 775.
[82] Certiorari denied, 137 S Ct 2290 (2017).

3.4 The Problem with Opacity

'Publicity', Jeremy Bentham wrote more than two centuries ago, 'is the very soul of justice. . . . It keeps the judge himself, while trying, under trial.'[83] Judicial decisions are the clearest example of an area in which the use of opaque AI systems should be limited, but even there we see 'algorithm creep'. As this chapter 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 with opacity.

The regulatory response to this opacity has been inconsistent. That is often the case with new technologies.[84] European efforts to restrain automatic processing clearly weigh the harmful social consequences more heavily than they are perceived in China. The US 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 chapter 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 can be improved by more robust testing and verification. Success is measured in the quality of those decisions, a cost–benefit analysis viewed through a utilitarian lens. Avoiding bias, by contrast, benefits 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.[85] 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.[86]

[83] Jeremy Bentham, 'Draught for the Organization of Judicial Establishments (1790)' in John Bowring (ed), *The Works of Jeremy Bentham* (William Tait 1843) vol 4, 285 at 316.
[84] See chapter seven, section 7.2.
[85] Some have gone further to suggest that these could be used for progressive purposes, a kind of 'algorithmic affirmative action': Anupam Chander, 'The Racist Algorithm?' (2017) 115 Michigan Law Review 1023, 1039–45.
[86] See chapter two, section 2.2.

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 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 AI 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 going on in a judge's head; it's a black box, too.'[87] As for Mr Loomis himself, he was released from Jackson Correctional Institution in August 2019 after serving his full six-year term. According to COMPAS, at least, there is a high risk that he will return.

[87] Jason Tashea, 'Risk-Assessment Algorithms Challenged in Bail, Sentencing and Parole Decisions', *American Bar Association Journal* (1 March 2017) (quoting Christine Remington).