6

Transparency

In July 2015, a group of hackers calling themselves the Impact Team broke into a Canadian company's website, stealing its user database and eight years of transaction records. A few weeks later they began posting online the personal information of more than 30 million customers. Data breaches are not uncommon, but the company in question was Ashley Madison, whose business model was based on arranging extramarital liaisons under the slogan 'Life is short. Have an affair.' The details posted included not only names and billing information but sexual preferences and fantasies.

The breach was initially greeted with a degree of schadenfreude: a bunch of adulterers were getting what they deserved. Yet as journalists pored over the data looking for celebrity gossip, a different news story developed. An online magazine known primarily for science fiction and tech scoops dug deeper and revealed that the frisson of scandal might have been misplaced. The vast majority of interactions on AshleyMadison.com were not between adulterous couples but between humans – almost all of whom were male – and bots. 'This isn't a debauched wonderland of men cheating on their wives,' Annalee Newitz wrote in *Gizmodo*. 'It's like a science fictional future where every woman on Earth is dead, and some Dilbert-like engineer has replaced them with badly designed robots.'[1]

The following month an unusual class action lawsuit was filed in Maryland, seeking compensation on the basis not of mishandling of data but of fraud. Specifically, Christopher Russell alleged – 'on behalf of himself and all others similarly situated' – that the site had deployed 'artificial intelligence "bots" and falsified user profiles to induce users to make purchases'.[2] He sought compensation for deceptive conduct in

[1] Annalee Newitz, 'Almost None of the Women in the Ashley Madison Database Ever Used the Site', *Gizmodo* (26 August 2015).

[2] *In re Ashley Madison Customer Data Security Breach Litigation*, 148 FSupp 3d 1378, 1380 (2015).

violation of consumer protection laws, and for unjust enrichment based on the company's bad faith conduct. Doubling down on the $100 he had spent purchasing credits on the site to chat with 'women', Russell (who had separated from his wife when he joined the site) claimed damages in excess of $5 million.

It is now almost three decades since the *New Yorker* published a cartoon that became one of the first Internet memes. A dog pauses from typing on a desktop computer and turns to another dog sitting nearby: 'On the Internet,' the first one explains, 'nobody knows you're a dog.' Questions of anonymity and pseudonymity have long bedevilled regulators of online behaviour. For many, the primary concern was identity theft, with neologisms like phishing coined to describe the theft of sensitive data through impersonation. The increasing sophistication of AI systems now means that many online processes – legitimate and not – are automated and handled through bots.

For human users, the ability to know whether they are interacting with another human is coming to be seen as a basic right.[3] This is a threshold question of transparency and not particularly controversial.[4] Larger questions of transparency relate to the increasing opacity of AI systems. As chapter three argued, this poses discrete challenges since it may allow inferior, impermissible, and illegitimate decisions to be made without the kind of scrutiny and accountability that would accompany human conduct of a similar nature.

The remedy is typically said to be transparency or 'explainability' – another neologism – with new areas of scholarship emerging on XAI and a novel 'right to explanation' thought to have been created by the EU in its GDPR. Such terms are often used imprecisely, however, in particular as between those versed in the technology and those versed in the law.

This chapter examines the extent to which transparency is desirable and possible in the context of AI systems, as well as how it is being defended in key jurisdictions. The first section unpacks the meaning of transparency, explainability, and related terms: what information can and should be made available, when, to whom, and at what cost. The second section then turns to how these concepts map onto advances in computer science and the field of XAI, as well as how any gaps may be filled. A third section considers how regulators, most prominently the

[3] White Paper on Artificial Intelligence (European Commission, COM(2020) 65 final, 19 February 2020) 20.

[4] In practice, however, it may not be so easily implemented. See below section 6.4, and chapter seven, section 7.4.

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

EU, have responded in operationalizing a legal right to certain forms of transparency, including (perhaps) a right to explanation.

The key finding is that there is a mismatch among these three discourses that over-emphasizes individualized after-the-fact explanations at the expense of addressing the systemic problems identified in chapter three. Episodic review can, in some circumstances, be *worse* than opacity because it gives the illusion of transparency. As naturally opaque AI systems become more prevalent, generating 'reasons' for a single decision may continue to be possible. But in the absence of systemic and proactive measures to ensure genuine transparency – or to make up for its absence – we risk losing the forest for the trees.

6.1 In Theory

Opacity was defined in chapter three as meaning the quality of being difficult to understand or explain. Transparency is used here to denote its opposite.[5] Similarly, the ends of transparency are in counterpoint to the dangers posed by opacity: it should improve the quality of decisions, deter or reveal impermissible decisions, and increase legitimacy and trust.[6] With regard to proprietary and complex opacity, existing tools enable regulators and courts to compel the disclosure of trade secrets or to recruit expert witnesses. The focus here is on naturally opaque AI systems that cannot be rendered meaningfully transparent without fundamental changes to the system itself.

'All models are wrong,' as George Box warned, 'but some are useful.'[7] The aphorism highlights a central challenge for explainability, which has emerged as the alternative to transparency. An explanation, in this context, means a description of how certain factors were used to reach a particular decision. In order to be useful, it must be comprehensible and enable an interested person to understand the extent to which specific inputs influenced the output. This includes what factors were used and whether changing one or more of them would have yielded a different

[5] See generally David Heald, 'Varieties of Transparency' in Christopher Hood and David Heald (eds), *Transparency: The Key to Better Governance?* (Oxford University Press 2006) 25. Among other things, Heald makes useful distinctions between event and process transparency, transparency in retrospect versus transparency in real time, and nominal as opposed to effective transparency.

[6] Cf Tal Z Zarsky, 'Transparent Predictions' [2013] University of Illinois Law Review 1503.

[7] George EP Box and Norman R Draper, *Empirical Model-Building and Response Surfaces* (Wiley 1987) 424.

result; it should also enable a comparison between decisions, revealing the reasons for the difference or similarity.[8]

A reasonable aim, perhaps, but it presents two problems. The first is that it necessarily requires simplification of the original system to make it comprehensible. The second is that it presumes that the purpose of an explanation is to help one individual understand a single decision. As we will see, that is only part of what explainability can mean – and only a fraction of what transparency should.

6.1.1 What?

Most writers distinguish between two ways of understanding the workings of an otherwise opaque AI system.[9] Early work on accountability of algorithms aspired to transparency in a general sense. Sometimes termed global or model-centric interpretability, it sought to disclose how an AI system functions. At one extreme, the entire code of the system might be published – fully transparent in one sense, but not particularly helpful if its workings are naturally opaque. More useful might be a description of the intentions behind the model, the training data that was used, performance metrics, and so on.[10] As such systems become more elaborate, however, the gap between representation and reality increases.

A second approach therefore turned to instance-based explanations, also termed local or subject-centric interpretability: understanding the factors influencing a particular decision. The emphasis is less on how the model functions than on why a particular decision was made in the way that it was.[11] It yielded different forms of explanation. Which factors were important, for example, in the outcome? Would variation in one of those factors have led to a different outcome?[12] This recalls one of the

[8] Finale Doshi-Velez and Mason Kortz, Accountability of AI under the Law: The Role of Explanation (Berkman Klein Center for Internet & Society, 2017) 2–3.

[9] See, eg, Christoph Molnar, *Interpretable Machine Learning* (Lulu 2019). Cf Riccardo Guidotti et al, 'A Survey of Methods for Explaining Black Box Models' (2018) arXiv 1802.01933v3, 16 (distinguishing among four possible approaches: explaining the model, explaining the outcome, inspecting the black box model internally, or providing a 'transparent solution').

[10] Lilian Edwards and Michael Veale, 'Slave to the Algorithm? Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking For' (2017) 16 Duke Law & Technology Review 18, 55–56. See generally David Brin, *The Transparent Society: Will Technology Force Us to Choose between Privacy and Freedom?* (Addison-Wesley 1998).

[11] Brent Mittelstadt, Chris Russell, and Sandra Wachter, 'Explaining Explanations in AI' (2018) arXiv 1811.01439v1, 2.

[12] This is sometimes termed counterfactual faithfulness.

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

benefits of AI systems, which is their ability to repeat decision-making processes while altering specific variables.[13] Importantly, it is possible to generate such an explanation without knowing the details of how the system reached the decision in question.[14]

The shift from general to specific was driven by utility. A global model merely approximating a naturally opaque AI system is unhelpful in understanding either how a system works or how a decision was reached. The more targeted explanation – the factors influencing the decision to deny a loan, whether the result would have been different if a higher salary had been reflected – at least gives affected persons more information and the possibility of changing their behaviour to achieve a different outcome. As we will see, however, these 'local' explanations can create problems of their own.

6.1.2 When?

A separate distinction is when the question of transparency should be considered. In the literature on regulation, two broad theories of oversight are known as 'police patrol' and 'fire alarm'. In the former, a sample of activities is investigated with the aim of detecting and remedying problematic behaviour and, through such surveillance, discouraging it. In the latter, a system is put in place where interested groups are empowered to raise an alarm and thus set in motion a response.[15] In the context of AI systems, the temporal question is typically broken down to 'before' and 'after' a decision is made. Neither maps neatly onto the oversight metaphor: naturally opaque systems may not reveal problems through *ex ante* sampling; reliance upon *ex post* alarms will generate a response only if a user knows that he or she has been harmed. Moreover, the 'before' stage may in fact be multiple stages: design of the model, selection of training data, validation, and so on.[16]

[13] See chapter three, section 3.2.2.

[14] Doshi-Velez and Kortz (n 8) 7.

[15] Mathew D McCubbins and Thomas Schwartz, 'Congressional Oversight Overlooked: Police Patrols versus Fire Alarms' (1984) 28 American Journal of Political Science 165, 166–76.

[16] Further complications arise in the case of machine learning algorithms that themselves change over time. See, eg, Mike Ananny and Kate Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2018) 20 New Media & Society 973, 982; Machine Learning Workflow (Google Cloud, 2020).

The shift to explainability reflects an acceptance that accountability, if it is to be sought at all, will be necessarily after-the-fact. But this gives rise to a pair of problems. The first is that it imposes a significant burden on users, many of whom will be unaware of adverse decisions, or unwilling or unable to challenge them. Support may be offered by requirements for audit trails in certain industries and failure accountability mechanisms analogous to the flight data recorders used to investigate aviation incidents. These will be discussed in the following section.

The second problem is that it discounts the value of regulation. In addition to the possibility of prohibiting certain forms of opaque decision-making completely, algorithmic impact assessments can be used to estimate the potential harms of automation. Periodic audits of sector-specific algorithms can also be used to detect bias without waiting for aggrieved individuals to step forward. Implementing such measures would be helped by tasking an institution with organizing the 'patrol'. These measures will also be considered below.[17]

6.1.3 To Whom?

The question of who can enforce transparency requirements or exercise rights of explanation also has two dimensions in the context of opaque decision-making.

The first is the threshold for invoking such powers and rights. In the context of transparency, the oversight role of regulators may bring with it the ability to demand access to information about how an AI system functions. If information cannot be extracted, the outsourcing of certain functions could be prohibited. Regulators in this context would include entities that monitor specific sectors – consumer credit, for example – or agencies like the police. They may also include new institutions, such as the 'algorithmic ombudsperson' considered in the next section. To the extent that rights are to be enforced by those entitled to seek an explanation of conduct by an AI system, requirements of standing may create the Catch-22 identified in chapter three: only those adversely affected by a decision have the right to bring an action, but in some cases no one will know about the adverse effects until an action has been brought.[18]

[17] See below section 6.2.2, and chapter eight, section 8.3.4.

[18] See chapter three, section 3.3.1. One approach, drawing on data protection law, would be to regard injuries in law as injuries in fact for the purposes of standing. See, eg, *Patel v Facebook, Inc*, 932 F 3d 1264 (9th Cir, 2019).

The second aspect concerns the knowledge or expertise that can be presumed on the part of a regulator or user. It is commonly said that information disclosed must be 'interpretable', for example, in the sense of being able to be understood by a human. But just any human? There is a significant difference between explaining machine learning processes to a computer scientist and explaining them to a lay person, but there is no agreed technical standard for comprehensibility by a human – even though that is precisely the point of explainability.[19] To the extent that only computer scientists are able to understand the work of their peers, putting technical experts in charge of accountability further runs the risk of regulatory capture.[20]

6.1.4 At What Cost?

A last consideration is that transparency is not free. As discussed in chapter three, even if a blanket prohibition on opaque AI systems were possible, it is not called for. Apart from anything else, a ban would mean that we forgo the many benefits that AI offers. Yet requiring that AI systems be 'transparent' also constrains innovation or introduces inefficiencies. Companies may be unwilling to expose trade secrets or invest the time and effort to develop sophisticated algorithms if they fear that these will be disclosed to competitors. Limiting the permissible number of variables in a model may render it more interpretable, but at the price of diminished accuracy.[21]

Insofar as explanations require responding to user complaints, processing those complaints also has a cost.[22] Proposals that counterfactual

[19] Cf Paul B de Laat, 'Algorithmic Decision-Making Based on Machine Learning from Big Data: Can Transparency Restore Accountability?' (2018) 31 Philosophy & Technology 525; Richard Tomsett et al, 'Interpretable to Whom? A Role-Based Model for Analyzing Interpretable Machine Learning Systems' (2018) arXiv 1806.07552; Danding Wang et al, 'Designing Theory-Driven User-Centric Explainable AI' (2019) CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems Paper No 601; Umang Bhatt et al, 'Explainable Machine Learning in Deployment' (2020) 1909.06342v4 arXiv.

[20] Leif Hancox-Li, 'Robustness in Machine Learning Explanations: Does It Matter?' (2020) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 640. See chapter eight, section 8.1.2.

[21] See, eg, AI in the UK: Ready, Willing, and Able? (House of Lords Select Committee on Artificial Intelligence, HL Paper 100, 2018), para 99. For a contrary view, arguing that the greater accuracy of complex models is often overstated, see Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1 Nature Machine Intelligence 206.

[22] See, eg, General Data Protection Regulation 2016/679 (GDPR) 2016 (EU), art 12(5) (allowing fees to be charged for manifestly unfounded or excessive requests for information).

explanations be generated for all automated decisions – positive and negative – would incur yet more costs.[23] Additional burdens may be associated with the information that is made public, ranging from the inadvertent disclosure of personal data to gaming of the system by users whose true motive is not to understand a decision but to manipulate it.

Some costs may be seen as investments in the legitimacy and integrity of AI systems, but efficiently allocating them requires a balancing of the potential harms against the impact of such measures. Too often, however, these conversations run in parallel or opposite directions, with rights-based advocates focused on the needs and interests of consumers even as technology races ahead.

6.2 In Practice

The practical matter of achieving either transparency or explainability of an AI system begins with the question of whether one has access to its inner workings. Transparency or explainability by design presumes access as well as an openness to prioritizing those qualities, even at the expense of functionality. Yet lacking access to the black box does not mean that explanations are impossible. Inputs and outputs offer windows into a model's performance and one AI system, it turns out, can be reasonably effective at extrapolating from the incomplete data of another.[24]

6.2.1 Methods

Transparency was originally sought through revealing those inner workings. In addition to disclosing source code in its entirety, this is sometimes considered at the level of components (decomposability)[25] or training algorithms (algorithmic transparency).[26] Naturally opaque systems have tested the limits of such approaches, with growing interest in

[23] See below section 6.2.1.

[24] See, eg, Wojciech Samek et al, 'Evaluating the Visualization of What a Deep Neural Network Has Learned' (2017) 28(11) IEEE Transactions on Neural Networks and Learning Systems 2660.

[25] See, eg, Grégoire Montavon et al, 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition' (2017) 65 Pattern Recognition 211.

[26] See, eg, Anupam Datta, Shayak Sen, and Yair Zick, 'Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems' (2016) IEEE Symposium on Security and Privacy (SP) 598. The related idea of inspectability means the ability to examine the logic and rules embedded in the system.

simulatability: developing a mechanistic understanding of models based on their performance – *what* they do in particular cases, rather than *how* they do it. The shift has meant that exogenous methods, which do not require access to the inner workings,[27] are more effective than might be presumed. Indeed, 'pedagogical', 'surrogate', and other 'model agnostic' approaches have seen significant advances in recent years.[28]

As indicated earlier, this is part of a larger move from transparency, in the sense of global interpretability, to explainability, in the sense of explaining specific decisions. Unhelpfully, the two are sometimes conflated. The Institute of Electrical and Electronics Engineers (IEEE) report on ethically aligned design, for example, lists transparency as one of eight general principles but defines it as meaning that the 'basis of a particular . . . decision should always be discoverable'.[29] The Asilomar Principles similarly limit transparency to 'failure transparency', meaning that an explanation will be required if an AI system causes harm.[30]

The shift was partly driven by what would be useful to users, especially unhappy users. It is also easier. AI system design can facilitate explainability after the fact. Some older systems that follow rules, decisions trees, or linear models can be written with automated explanations built in.[31] The IEEE acknowledges the limitations of predicting what more advanced AI systems will do and therefore advocates traceability as a means of mitigating any harm.[32] Others have sought to propose standards for recording model performance characteristics and training data.[33]

But this is not the same as transparency. A 2017 report by the US Defense Advanced Research Projects Agency (DARPA) similarly stated that the aims of XAI are enabling human users to 'understand, appropriately trust, and effectively manage the emerging generation of artificially

[27] See, eg, Ashley Deeks, 'The Judicial Demand for Explainable Artificial Intelligence' (2019) 119 Columbia Law Review 1829, 1835.

[28] Alejandro Barredo Arrieta et al, 'Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities, and Challenges Toward Responsible AI' (2020) 58 Information Fusion 82, 82–84. One of the better known is Local Interpretable Model-Agnostic Explanations (LIME) and its variants. Others include Bayesian Rule Lists (BRL) and Shapley Additive Explanations (SHAP).

[29] Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems (IEEE, 2019) 27.

[30] Asilomar AI Principles (Future of Life Institute, 6 January 2017).

[31] Alberto Blanco-Justicia et al, 'Machine Learning Explainability via Microaggregation and Shallow Decision Trees' (2020) 194 Knowledge-Based Systems 105532, 1–2.

[32] Ethically Aligned Design (n 29) 137.

[33] See below section 6.2.2(b).

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

intelligent partners'.[34] Understanding and trust are important, though the focus on individual users is made evident in measurements of explanation effectiveness such as 'user satisfaction'.

Other studies have attempted to measure XAI, ascribing quantitative values to its 'goodness', usefulness and satisfaction to users, improvement of their mental models, as well as the impact of explanations on the performance of the model and on the trust and reliance of the audience.[35] One of the more prominent forms of explanation is termed counterfactual – meaning that the explanation seeks to highlight how alternative outcomes might have been reached with different input variables.[36] Analogous to the 'principal reason' explanation required in US credit laws, the intention is to provide users with actionable guidance – for example, how changed financial circumstances might have enabled them to get a loan or lower interest rate. Advantages include that such explanations can now be generated without human intervention and without needing to disclose the underlying model. Limitations are that it works best on binary outcomes, presumes that the model remains stable over time, and relies on the changed values mapping onto real-world actions while other factors remain constant.[37]

These are all important developments, but, from the perspective of regulation, different audiences and distinct interests arise. The focus on explaining decisions that depart from what users want or expect captures only a fraction of the decisions made, while the rest may be accepted because users do not complain and automation bias fills in any gaps.[38] Yet if the aims of transparency are to improve decision quality, prevent or punish impermissible decisions, and increase legitimacy, then more is needed. Technical solutions must be supplemented with regulatory ones

[34] David Gunning, Explainable Artificial Intelligence (XAI), Program Update (Defense Advanced Research Projects Agency (DARPA), DARPA/I2O, November 2017) 7.

[35] Robert R Hoffman et al, 'Metrics for Explainable AI: Challenges and Prospects' (2019) arXiv 1812.04608v2; Sina Mohseni, Niloofar Zarei, and Eric D Ragan, 'A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems' (2020) arXiv 1811.11839v4.

[36] See, eg, Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan, 'Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations' (2020) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 607.

[37] Solon Barocas, Andrew D Selbst, and Manish Raghavan, 'The Hidden Assumptions behind Counterfactual Explanations and Principal Reasons' (2020) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 80.

[38] See chapter three, section 3.1.

and local explanations must be complemented by some measure of global transparency.[39]

6.2.2 Tools

Three promising regulatory tools will be considered here: algorithmic impact assessments, algorithmic audits, and an AI ombudsperson. The present discussion focuses on the limited context of transparency; broader institutional possibilities for regulating AI will be discussed in chapter eight.

(a) Algorithmic Impact Assessments

Algorithmic impact assessments have emerged as a specific application of data protection (or privacy) impact assessments, which were in turn based on environmental impact assessments. This genealogy is significant in two ways. First, the analogy with impact on the environment helpfully links assessments with existing policies and practices: a study, prior to committing to a project, of its likely consequences in an area of sensitivity. In the case of environmental impact, an evaluation of costs and benefits may conclude that a development should not proceed, or that safeguards should be taken to mitigate the harmful effects of pollution, disruption of wildlife, and so on. This technique dates back at least to 1970, when it was first introduced into US law.[40]

Privacy impact assessments came considerably later, with legislation in New Zealand in 1993,[41] soon followed by Canada, Australia, and the United States.[42] The EU included in its 1995 Data Protection Directive a requirement that member states make a determination of the risks to

[39] The field of XAI moves quickly and the present work does not attempt to do justice to the computer science literature. See, for example, the annual ACM Conference on Fairness, Accountability, and Transparency, available at https://facctconference.org. The conference was initially known by the acronym FAT, but in 2020 changed this to FAccT. A related field of algorithmic fairness is also beyond the scope of the present study. See, eg, Pak-Hang Wong, 'Democratizing Algorithmic Fairness' (2020) 33 Philosophy & Technology 225.

[40] National Environmental Policy Act 1970 (US). See generally Neil Craik, *The International Law of Environmental Impact Assessment: Process, Substance and Integration* (Cambridge University Press 2008).

[41] Privacy Act 1993 (NZ), s 105.

[42] See, eg, Electronic Government Act 2002 (US). Cf Federal Privacy Act 1974 (US) 5 USC § 552a(r) requiring public agencies changing record systems to allow for evaluation of the effect on privacy rights.

rights and freedoms of certain activities,[43] but a formal data protection impact assessment (DPIA) came only with the GDPR in 2016. That GDPR requirement, linked with other provisions on automated processing, was the next stepping stone to algorithmic impact assessments and the second reason to note their pedigree. For it is limited to 'the impact of the envisaged processing operations on the protection of personal data'.[44] As we have seen, the negative consequences of opaque AI systems may include – but are certainly not limited to – the impact on personal data.

In theory, algorithmic impact assessments should enable people to know which systems affect their lives, increase the quality of decisions, and ensure greater accountability by enabling experts as well as affected individuals to review automated processes.[45] Unlike the narrower right to explanation, the purpose of an algorithmic impact assessment is to ensure that documentation is available before decisions are made.[46] An ideal process would see an organization make public details about each AI system it intends to use and undertake an assessment of potential harms, as well as the means of addressing those harms on an ongoing basis. It should allow for a comment period during which individuals potentially affected could challenge either the harms that have been flagged or the proposed response.[47]

Assessments work best when undertaken at the level of a specific project rather than an organization, are done in advance rather than in retrospect, take a broad approach in terms of the stakeholder interests as well as the norms considered, and focus on solving rather than merely highlighting problems.[48] In practice, however, the track record of DPIAs shows that they focus on data quality and security rather than the broader social and legal impacts.[49]

[43] 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 20.

[44] GDPR, art 35(1). See below section 6.3.1.

[45] Dillon Reisman et al, Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability (AI Now, April 2018) 5.

[46] Andrew D Selbst, 'Disparate Impact in Big Data Policing' (2017) 52 Georgia Law Review 109, 169–93.

[47] Reisman et al (n 45) 9–10.

[48] Margot E Kaminski and Gianclaudio Malgieri, 'Multi-layered Explanations from Algorithmic Impact Assessments in the GDPR' (2020) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 68, 71.

[49] Alessandro Mantelero, 'AI and Big Data: A Blueprint for a Human Rights, Social, and Ethical Impact Assessment' (2018) 34 Computer Law & Security Review 754, 761–62.

A further limitation is that they are often voluntary or, as in the case of the GDPR, give significant latitude to organizations.[50] It may be impractical to require a full assessment for every AI system used by every organization. As argued in chapter three, however, this could be managed in a tiered fashion. There is a strong argument that inherently governmental functions should not be outsourced at all, for example: properly designed impact assessments could help determine whether an opaque AI system should be deployed, and with what safeguards.

(b) Algorithmic Audits

Audits, in particular external audits, are commonly used to improve processes and guard against wrongdoing. They are also intended to verify whether information disclosed – financial statements, for example – reflects a true and fair view of a company's financial position. In the case of opaque AI systems, audits can be used to determine whether an algorithm is behaving in the manner intended, and whether it is prone to impermissible bias.[51] Audit logs provide a useful record that can be reviewed to see the provenance of training data or the aggregate effect of a model on a user population, but even more important is the ability to impersonate new users and systematically test for biased outcomes such as those discussed in chapter three.[52]

Even so, it may be challenging to define what factors amount to impermissible bias and how to test for them. Obvious candidates are those protected by national anti-discrimination laws – sex/gender, race, age, religion, disability, and so on.[53] Searching for bias may pose difficulties if there is no baseline against which to measure, however. 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.[54] Even good faith efforts to use

[50] See below section 6.3.1.

[51] The IEEE standard for software development defines an audit as 'an independent evaluation of conformance of software products and processes to applicable regulations, standards, guidelines, plans, specifications, and procedures'. IEEE Standard for Software Reviews and Audits (IEEE, Standard 1028-2008, 2008).

[52] See chapter three, section 3.2.

[53] See generally Tarunabh Khaitan, *A Theory of Discrimination Law* (Oxford University Press 2015); Nina Grgić-Hlača et al, The Case for Process Fairness in Learning: Feature Selection for Fair Decision Making (Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems (NIPS), 2016).

[54] Karen Hao, 'This Is How AI Bias Really Happens – and Why It's So Hard to Fix', *MIT Technology Review* (4 February 2019).

algorithms to combat bias will fail if they are unable to take account of social context. 'Fairness', for example, is a property not of a technical system but of the society within which that system functions.[55]

A further difficulty is that algorithmic audits typically treat the AI system being examined as a black box, limiting the ability to infer causes from different testing outcomes or to determine whether further variations in inputs would have led to different (and potentially problematic) outputs.[56] A more promising approach is to conduct audits at each stage of the development process, especially at the early stages of model development, in conjunction with the development of a risk register.[57] This may run against the culture of technological innovation – audits are necessarily methodical, boring, and slow – but internal audits at defined stages and diligent record-keeping throughout may be the only way to identify and prevent certain impermissible decisions before – or after – they are made.[58]

Two recent standards may be useful in creating such documentation. Model cards include information about how a model was built, the assumptions made during its development, and the kinds of behaviour that might be experienced by different demographic groups.[59] Datasheets for machine learning datasets draw an analogy with documentation of hardware in the electronics industry and propose that every dataset be accompanied by a datasheet that documents its motivation, composition, collection process, recommended uses, and so on.[60]

(c) AI Ombudsperson

The fundamental problem of AI opacity is that one doesn't know what one doesn't know. Most existing accountability regimes rely on aggrieved individuals initiating proceedings against the developer or owner of an

[55] Richard Berk, *Machine Learning Risk Assessments in Criminal Justice Settings* (Springer 2019) 115–30; Andrew D Selbst et al, 'Fairness and Abstraction in Sociotechnical Systems' (2019) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 59. Cf Ifeoma Ajunwa, 'The Paradox of Automation as Anti-bias Intervention' (2020) 41 Cardozo Law Review 1671.

[56] Cf Joshua A Kroll et al, 'Accountable Algorithms' (2017) 165 University of Pennsylvania Law Review 633, 661.

[57] See Fiona D Patterson and Kevin Neailey, 'A Risk Register Database System to Aid the Management of Project Risk' (2002) 20 International Journal of Project Management 365.

[58] Inioluwa Deborah Raji et al, 'Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing' (2020) 2001.00973v1 arXiv.

[59] Margaret Mitchell et al, 'Model Cards for Model Reporting' (2019) ACM Conference on Fairness, Accountability, and Transparency (FAT*) 220.

[60] Timnit Gebru et al, 'Datasheets for Datasets' (2020) arXiv 1803.09010v7.

opaque AI system, with all the barriers to success noted earlier.[61] Impact assessments before, coupled with internal and external audits during and after deployment will address some of the concerns about inferior, impermissible, and illegitimate decisions. But, in many jurisdictions, it would be helpful to have an institution able to investigate complaints that do not easily fit within existing causes of action and to represent the public interest with respect to systemic issues.

An ombudsperson (or ombudsman) is one such institution and has been mooted periodically in the context of algorithms or AI more generally. If created, its mandate would extend significantly beyond questions of transparency or explainability, but it is mentioned here because the relative flexibility could enable proportionate responses to situations in which information is limited due to uncertainty concerning the underlying technology. It will be discussed more fully in chapter eight.[62]

6.3 In Law

As in many areas of AI regulation, technology has raced ahead of law on questions of opacity and transparency. Some jurisdictions have embraced this. Singapore, for example, has adopted a technology-neutral model framework instead of legislation; even that non-binding document acknowledges that 'perfect explainability, transparency and fairness are impossible'.[63] Jurisdictions that have legislated (or tried to) each face the dilemma of constraining innovation or finding themselves unable to contain its undesirable consequences.[64] As the experience of the EU shows, they may also find that compromise language intended to square that circle introduces uncertainties of its own.

6.3.1 *An EU Right to Explanation?*

While the clock counted down towards the entry into force of the GDPR on 25 May 2018, a curiously heated debate spread across the pages of journals more accustomed to staid academic commentary. Text that had taken four years to negotiate had fundamentally changed the

[61] See above section 6.1.3.
[62] See chapter eight, section 8.3.4.
[63] Model Artificial Intelligence Governance Framework (2nd edn) (Personal Data Protection Commission, 2020) 15.
[64] See chapter seven, section 7.2.

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

transparency landscape by creating a new 'right to explanation'.[65] No it had not, came the counterattack, claiming that the GDPR had been fundamentally misread.[66]

Unsurprisingly, both sides were oversimplifications – enabled by an apparent disconnect between the GDPR's recitals and its text. Non-binding Recital 71 states that individuals should have 'the right ... to obtain an explanation of [a] decision reached' solely through automated processing.[67] Article 22 imposes limits on when processing is permissible and has been discussed in chapter two.[68] In terms of explanations, however, it is silent. An amendment had been proposed that would have included among the 'suitable measures ... the right to obtain human assessment *and an explanation of the decision reached*'.[69] That language was dropped from the GDPR, however, leading some to conclude that the 'right to explanation' had been considered and abandoned. This greatly overstated the case, also ignoring the fact that Article 15 includes a right to obtain information about the existence of automated decision-making and 'meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing' for a person.[70]

The argument degenerated into a semantic dispute over the difference between a 'right to explanation' and a 'right to ... meaningful information',[71] but it appears to have been settled by the European Data Protection Board. Previously known as the Article 29 Working Party, its guidelines on implementation of the GDPR provide that 'meaningful information' need not include a complex explanation of the algorithm or disclosure of the full algorithm but should be

[65] Bryce Goodman and Seth Flaxman, 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"' (2017) 38(3) AI Magazine 50.

[66] 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' (2017) 7 International Data Privacy Law 76.

[67] GDPR, Recital 71.

[68] See chapter two, section 2.3.

[69] Report of the Committee on Civil Liberties, Justice and Home Affairs on the proposal for a regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (European Parliament, COM(2012) 0011–C7-0025/2012–2012/0011(COD), 2013) (emphasis added).

[70] GDPR, art 15(1)(h).

[71] The German text of the GDPR uses the phrase '*aussagekräftige Informationen*', which is close to 'meaningful information', while the French ('*informations utiles*') and the Dutch ('*nuttige informative*') might be translated as 'useful information'.

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

'sufficiently comprehensive for the data subject to understand the reasons for the decision'.[72]

There are, nonetheless, significant loopholes. The relevant provisions apply only to decisions based 'solely on automated processing' that produce legal or similar effects.[73] Rights of access are also to be interpreted in a manner that does not 'adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software'.[74]

The GDPR also provides for a DPIA that may appear to overlap significantly with the algorithmic impact assessment proposed earlier.[75] Article 35 provides that an organization must assess a proposed system, including its necessity and proportionality in relation to stated purposes; the assessment must also cover the risks to the rights and freedoms of affected individuals, as well as measures to address those risks.

Again, however, there are significant limitations. The threshold for requiring a DPIA is initially said to be where there is a 'high risk to the rights and freedoms of natural persons', though this is later defined as including 'systematic and extensive evaluation of personal aspects' of natural persons leading to decisions with legal or comparable effects.[76] As one group of scholars wryly noted, demonstrating that a DPIA is not necessary may well itself require a DPIA.[77] Another shortcoming, for present purposes at least, is that it focuses on the protection of personal data. This is important but is hardly the only concern associated with the operation of opaque algorithms. In addition, the EU DPIA provides for only limited consultation – internally with a data protection officer and with data subjects themselves only 'where appropriate' and 'without prejudice to the protection of commercial or public interests or the security of processing operations'.[78] An earlier proposal that consultation be mandatory was dropped as it was thought to impose a disproportionate

[72] 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) 25. Cf Profiling and Automated Decision-Making (Information Commissioner's Office, 2017) 15.

[73] GDPR, art 22(1).

[74] Ibid, Recital 63.

[75] See above section 6.2.2(a).

[76] GDPR, art 35(1), (3).

[77] 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' (2019) 34 Berkeley Technology Law Journal 145, 176.

[78] GDPR, art 35(2), (9).

burden on data controllers. Similarly, an organization is required to consult the relevant data protection authority in its jurisdiction only if its own assessment concludes that there is a high risk in the absence of measures to mitigate it.[79] There is no requirement for the assessment to be made public.

Despite all these reservations, it is possible that the GDPR will have an impact. In January 2019, Google was fined €50 million by France's data protection agency, the *Commission nationale de l'informatique et des libertés* (CNIL) – the largest penalty imposed under the GDPR up to that point. The breaches by Google included its failure to provide information concerning the use of personal data in providing targeted advertising on its Android devices, leaving users unable 'to sufficiently understand the particular consequences of the processing for them'.[80] An appeal to the *Conseil d'État* [Council of State] was dismissed in June 2020, affirming that the relevant information was not presented in a sufficiently clear and distinct manner for the user's consent to be validly obtained.[81]

Though the case turned on transparency, the barriers to understanding had less to do with neural networks than with legalese. Explanations of how Google used personal data were spread across multiple documents that were vague and difficult to access, sometimes requiring up to five or six actions to find them. This amounted to a violation of the GDPR obligation to provide information about the collection and use of personal data in a 'clear' and 'intelligible' manner,[82] rather than challenging the use of an opaque AI system as such. Nonetheless, the decision is still noteworthy for the fact that it was initiated by two not-for-profit organizations representing a class of just under ten thousand users. Though no specific harm was alleged, this was deemed sufficient for the CNIL to commence an investigation, which took place only online. It is possible that future class actions will go after the use of opaque algorithms, though it is far simpler to show the inadequacy of language governing consent

[79] Ibid, art 36.

[80] Deliberation of the Restricted Committee Pronouncing a Financial Sanction Against Google LLC (CNIL, SAN-2019-001, 21 January 2019), para 111.

[81] RGPD: le Conseil d'État rejette le recours dirigé contre la sanction de 50 millions d'euros infligée à Google par la CNIL [GDPR: The Council of State Rejects the Appeal against the Sanction of 50 Million Euros Imposed on Google by the CNIL] (Conseil d'État, 19 June 2020).

[82] GDPR, art 12(1).

than to demonstrate opacity in how personal data so collected is being used.

6.3.2 *Council of Europe Convention 108*

The Council of Europe's Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data was first adopted in 1981 and entered into force in 1985. Aside from 1999 measures opening it up to the EU, the most important amendments were adopted in 2018 – a week before the GDPR entered into force – and sought to address the 'new challenges' posed by AI systems.[83]

Among other changes, these introduced new obligations of transparency in the sense of both the identity of data controllers as well as the legal basis and purpose of any processing. This was to be enforced by users having the right 'to obtain, on request, knowledge of the reasoning underlying data processing where the results of such processing are applied to him or her'.[84] As the explanatory report makes clear, however, this amendment presumes that data controllers have this information at their disposal and should make it available to data subjects.[85] It is therefore of limited application here.

6.3.3 *France*

In 2016, France passed its own *République numérique* [Digital Republic] law. Limited to administrative bodies, this creates a right to request information about algorithmic decisions, including the rules and main characteristics of the algorithm.[86] 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'.[87] This last

[83] Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (ETS No 108), done at Elsinore, Denmark, 17–18 May 2018.

[84] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Convention 108), done at Strasbourg, 29 January 1981, ETS No 108, in force 1 October 1985, art 9(1)(c).

[85] Explanatory Report to the Protocol amending the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data (Council of Europe, 10 October 2018), para 77.

[86] Loi no 2016-1321 du 7 octobre 2016 pour une République numérique 2016 (France), art 4.

[87] 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

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.[88]

6.3.4 *United States*

In April 2019, bills proposing a new Algorithmic Accountability Act were introduced in the US Senate and House. Driven by revelations of bias – the press release cited accusations that Facebook violated the Fair Housing Act by allowing advertisers to discriminate based on race, religion, and disability status – the proposed law would require impact assessments for 'high risk' automated decision-making systems concerning their 'accuracy, fairness, bias, discrimination, privacy, and security'. Limited to entities with revenue in excess of $50m or holding data of more than a million customers, it would not create a private right of action or operate extraterritorially. Enforcement would be through the US Federal Trade Commission (FTC) or state attorneys-general.[89]

Though unlikely to become law, the draft legislation is of interest for at least two reasons. First, it would address regulation of AI generally, rather than being sector-specific – something that has undermined the coherence of US privacy and data protection laws over the decades.[90] Consideration of discrimination would shift from enforcement based on a patchwork of existing laws to prevention or mitigation based on a single law. Secondly, its scope more properly covers how algorithms are developed and used. The proposed impact assessment encompasses the system itself as well as its development process, including its design and training data, though there is no requirement that the findings be made public. Unlike the GDPR, automated decision-making is defined as a computational process that makes a decision 'or facilitates human

(France), art 1. See also 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, 48–49.

[88] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR' (2018) 31 Harvard Journal of Law & Technology 841, 842–43.

[89] Algorithmic Accountability Act of 2019, S 1108, HR 2231, 116th Congress 2019 (US).

[90] Simon Chesterman, *One Nation under Surveillance: A New Social Contract to Defend Freedom without Sacrificing Liberty* (Oxford University Press 2011) 244.

decision making', avoiding the problem of confining protections to decisions based 'solely on automated processing'.[91]

The press release included the soundbite 'Algorithms shouldn't have an exemption from our anti-discrimination laws.' This oversimplified the regulatory challenge of addressing algorithmic bias – algorithms are not 'exempt' from the law. But it is true that their opacity makes it more difficult to discover or remedy discriminatory behaviour. If the legislation or something like it is passed, it would make it easier to do both.

6.3.5 Canada

In April 2020, Canada's Directive on Automated Decision-Making came into force. Covering most Federal administrative decisions, it requires that an algorithmic impact assessment be carried out prior to deploying 'any technology that either assists or replaces the judgment of human decision-makers'.[92] The impact assessment follows a standard form and must be completed prior to production.

The breadth of coverage is important, but of particular interest is the sliding scale of transparency requirements, based on the level of potential harm. Where decisions will likely have 'little to no impact' on individuals or communities, no notice is required. For decisions with 'moderate' impact, a plain language notice must be published on the programme or service's website. Where 'high' or 'very high' impact is anticipated, the website must also include a description of how the components work, how it supports the decision, the results of any reviews or audits, and a description of the training data (or a link to the anonymized data itself, if publicly available). A separate provision covers explanations of decisions. Those with minimal impact need provide only a 'frequently asked questions' section on a website. Moderate impact decisions should offer 'meaningful explanation' on request, while high and very high impact decisions that deny a benefit or service should include the explanation with the decision itself.[93]

Though limited to the public sector, the Canadian directive is one of the most progressive yet adopted. Government agencies had earlier been urged to use open source software; the directive adds a presumption that the custom source code of a system owned by the Canadian government

[91] GDPR, art 22(1).
[92] Directive on Automated Decision-Making 2019 (Canada), Appendix A.
[93] Ibid, s 6.2.

should also be released, subject to prescribed exceptions for classified and other data. It is too early to evaluate implementation of the directive, but some concerns have already been expressed about using it in immigration decisions in place of more formal, enforced standards.[94]

6.3.6 *Other Jurisdictions*

Other jurisdictions have considered ways to preserve or encourage transparency while taking advantage of AI, though most have remained in the realm of voluntary principles comparable to Singapore's Model Framework, discussed earlier.[95] In Australia, for example, the Federal government published a set of AI Ethics Principles in November 2019, among other things stating that people should be able to 'know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them'.[96] Some governments have followed the Canadian lead in exploring tighter restrictions for public sector processes. In July 2020, New Zealand published its 'Algorithm Charter', under which government agencies promise to 'clearly [explain] how significant decisions are informed by algorithms'. A draft had included that agencies would also state 'who is responsible for automated decisions', but this was dropped from the final text.[97]

For its part, China's Ministry of Science and Technology also adopted principles for AI governance in 2019 with the aim of ensuring that AI systems are safe, controllable, and reliable. The eight principles overlap somewhat with comparable frameworks elsewhere,[98] but transparency is not high among them. Fairness is to be promoted and discrimination 'eliminated', but transparency and interpretability are targeted for 'continuous improvement', while 'gradually achieving' auditability.[99]

[94] Fenwick McKelvey and Margaret MacDonald, 'Artificial Intelligence Policy Innovations at the Canadian Federal Government' (2019) 44(2) Canadian Journal of Communication 43, 46.

[95] See above n 63.

[96] AI Ethics Principles (Department of Industry, Science, Energy and Resources, November 2019). Cf the Commonwealth Scientific and Industrial Research Organisation (CSIRO) discussion paper earlier in 2019, which called for a broad interpretation of transparency. People should be informed 'when an algorithm is being used that impacts them and they should be provided with information about what information the algorithm uses to make decisions'. D Dawson et al, Artificial Intelligence: Australia's Ethics Framework (Data61 CSIRO, 2019) 6–7.

[97] Algorithm Charter for Aotearoa New Zealand (Department of Internal Affairs, July 2020).

[98] See chapter seven, introduction.

[99] 新一代人工智能治理原则——发展负责任的人工智能 [The New Generation of Artificial Intelligence Governance Principles – the Development of Responsible Artificial

https://doi.org/10.1017/9781009047081.008 Published online by Cambridge University Press

6.4 The Limits of Transparency

Transparency is a means, not an end. Its purpose is, in part, to avoid or limit the risks of opacity discussed in chapter three: inferior, impermissible, and illegitimate decisions. But transparency also builds trust. That is routinely acknowledged to be one of the major barriers to adoption and acceptance of new technologies in general and AI in particular.[100] The shift in focus that this chapter describes – from transparency to explainability – acknowledges individualized concerns about the use of AI and the practical challenges posed by natural opacity. While individual explanations may help correct some inferior decisions or reveal impermissible ones, however, this depends on affected users knowing that they have been harmed and being in a position to complain about it. Such explanations do not address the illegitimacy of a decision where reasons should precede rather than follow the making of it.

Even with the best of intentions and resources, it is important to be realistic about the limitations of transparency – indeed, openness about those limitations may be the most important form of transparency.[101] Clearly, transparency is not a panacea. But sometimes it is a distraction and sometimes it is undesirable.

As a distraction, illusory transparency can be worse than opacity. This illusion, sometimes termed the transparency fallacy, takes two forms. First, much as some governments demonstrate their commitment to 'openness' by burying constituents in unstructured records, the provision of vast amounts of data or source code may be transparency in form only. Secondly, a theoretical individual right to explanation that cannot in practice be exercised deflects criticism without providing a genuine remedy. Even if it is understood, in the absence of the possibility to use information to bring about systemic change, it will not help achieve meaningful accountability. Much as the theoretical consent of users has long provided the fig leaf for data protection law, the illusion of transparency could give false comfort to those seeking to hold AI systems to account.

Intelligence] (Ministry of Science and Technology, 17 June 2019), paras 2 ('消除 . . . 歧视'), 5 ('人工智能系统应不断提升透明性, 可解释性 . . . 逐步实现可审核').

[100] See, eg, Robin C Feldman, Ehrik Aldana, and Kara Stein, 'Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know' (2019) 30 Stanford Law & Policy Review 399.

[101] Cf Karl de Fine Licht and Jenny de Fine Licht, 'Artificial Intelligence, Transparency, and Public Decision-Making: Why Explanations Are Key When Trying to Produce Perceived Legitimacy' (2020) 35 AI & Society 917.

In some cases, though, transparency may be undesirable. Systems intended to maintain security or prevent fraud or other wrongdoing should preserve sufficient opacity to carry out their functions.[102] Disclosing details of algorithms – whom to screen more thoroughly at checkpoints, for example – might uncover bias but also enable manipulation.[103] If datasets are made public, the personal data that forms the basis for some decisions will be exposed. Even without the dataset, some model explanations can be exploited to reveal the underlying training data.[104] It is no coincidence that many legislative efforts at requiring transparency are found in data protection law.

A different critique of transparency and explainability is that we sometimes ask them to do too much. Calls for transparency on the part of AI systems often start from questionable assumptions about human decision-making – contrasting algorithmic processing, for example, with 'traditional decision-making, where human decision-makers can *in principle* articulate their rationale when queried, limited only by their desire and capacity to give an explanation, and the questioner's capacity to understand it'.[105] The 'in principle' is doing a lot of work here, as the process by which humans actually make decisions is known to be inextricably tied to intuition, hunches, personal impressions – with a layer of after-the-fact ratiocination.[106] When we require a human decision-maker to give reasons, we do not ask them to undergo functional magnetic resonance imaging in order to understand the cognitive process by which a decision was actually reached.

Language does not always help here. When considering explanations of different phenomena, we think of volitional human behaviour in terms of reasons rather than causes. When explaining a human decision, it would be odd to present the *cause* of a particular choice. Though we might say that new shoes cause us to walk in a particular way, we would not say that their discounted price 'caused' us to buy them. In the physical world, the reverse is true: we would not normally speak of the *reason* a fire

[102] Jenna Burrell, 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms' (2016) 3(1) Big Data & Society 4.
[103] Anupam Chander, 'The Racist Algorithm?' (2017) 115 Michigan Law Review 1023, 1034.
[104] Reza Shokri, Martin Strobel, and Yair Zick, 'On the Privacy Risks of Model Explanations' (2020) arXiv 1907.00164v5.
[105] Brent Daniel Mittelstadt et al, 'The Ethics of Algorithms: Mapping the Debate' (2016) 3 (2) Big Data & Society 7 (emphasis added).
[106] John Zerilli et al, 'Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?' (2019) 32 Philosophy & Technology 661, 665–68.

started, except perhaps as the prelude to an explanation about a cause.[107] The language of 'reasons' presumes a degree of subjectivity and rationality on the part of an actor: they belong to that actor in a way that causes do not.[108] In the case of computers, then, the demand for reasons suggests another form of anthropomorphism. From what *caused* the computer to do *x* (shut down, say, or catch fire), we slide into what were the computer's *reasons* for doing *y* (deny me a loan, suggest that I watch a particular movie, and so on).

Transparency in AI systems is sought not for its own sake but for purposes similar to why it is sought in human decisions. The methods of achieving it are distinct, however. Sometimes that is beneficial – such as when we can test for bias by running multiple simulations without worrying that an AI system will become defensive and dissemble. Sometimes it is challenging. Useful explanations offer contrasts – not merely why *x* happened but why *x* rather than *y* – and they are selective in the sense that they prioritize relevance and context over completeness. Human explanations emphasize factors influencing a decision rather than raw probabilities and are expressed in a manner that is tailored to the world views of the parties concerned. None of this is easy for an AI system. And sometimes the difference between AI and human explanations can be misleading. Where there is discretion to be exercised, for example, it can be artificial to ascribe reasons.

This is particularly the case for certain public decisions, where the legitimacy of the outcome is tied to the identity of the decision-maker.[109] Many arguments warning of opaque AI systems determining the fate of humans conjure a dystopian world without explanations, epitomized by Franz Kafka's *The Trial*.[110] A man, known only as Josef K, is arrested and prosecuted by unknown agents for an unknown crime; attempts to understand or escape his absurd ordeal are fruitless. The metaphor is a compelling one. But it is flawed. For the power of *The Trial* is not that there are hidden explanations being withheld from the hapless Josef K but that there is no logic to his predicament at all.

[107] For example: 'The reason that the fire started was because . . .' in which the opening of the sentence is redundant.

[108] Tim Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2019) 267 Artificial Intelligence 1, 16.

[109] See chapter three, section 3.3.

[110] See, eg, Andrew Selbst and Solon Barocas, 'The Intuitive Appeal of Explainable Machines' (2018) 87 Fordham Law Review 1085, 1118; Daniel J Solove, 'Privacy and Power: Computer Databases and Metaphors for Information Privacy' (2001) 53 Stanford Law Review 1393, 1419–23.

The increased role of AI systems in making or assisting decisions will, in a great many cases, optimize outcomes; individuals who are denied a service or adversely affected by such decisions are entitled to an explanation. But avoiding inferior, impermissible, or illegitimate decisions requires more than this. Impact assessments before, audits during, and an independent advocate after those decisions will lead to better decisions as well as increasing trust in those decisions.

The starting point is transparency as to the involvement of AI systems in the first place. At present this is, for the most part, a simple yes or no question. Increasingly, however, AI-assisted decision-making will blend human and machine. Some chatbots start on automatic (human-out-of-the-loop) for basic queries, moving through suggested responses that are vetted by a human (over-the-loop), escalating up to direct contact with a person (in-the-loop) for unusual or more complex interactions.[111] Though decisions based 'solely on automated processing' are going to increase, 'machine-assisted' decisions will skyrocket. Much as passengers in autonomous vehicles need clarity as to who is meant to be holding the wheel, humans interacting with AI systems should be aware of with whom – or with what – they are dealing.

This goes both ways. To guard against overuse or malicious attacks, many websites now utilize challenge tests such as CAPTCHA. These function as a kind of reverse Turing Test, with computers requiring proof that a user is not another computer.[112]

A year after the Ashley Madison scandal, the parent company had a new name and chief executive, while the company itself adopted a new slogan: 'Find your moment.' In place of a wedding ring and a woman putting a finger to her lips, a ruby logo was deemed more 'multi-faceted' and relevant to a wider user-base. In addition to the class action lawsuit, it was reported that the US FTC was investigating the company's use of bots to chat with paying male customers.

Within another twelve months, however, the 'have an affair' language returned, as did the company's focus on connecting adulterous couples.

[111] See, eg, Pavel Kucherbaev, Alessandro Bozzon, and Geert-Jan Houben, 'Human-Aided Bots' (2018) 22(6) IEEE Internet Computing 36.

[112] CAPTCHA is a contrived acronym standing for Completely Automated Public Turing test to tell Computers and Humans Apart. See Luis von Ahn et al, 'CAPTCHA: Using Hard AI Problems for Security' in Eli Biham (ed), *Advances in Cryptology – EUROCRYPT 2003* (Springer 2003) 294; Henry S Baird, Allison L Coates, and Richard J Fateman, 'PessimalPrint: A Reverse Turing Test' (2003) 5 International Journal on Document Analysis and Recognition 158.

In an interview with the *New York Post*, its Vice President for Communications said that the summer of 2015 had led to 'unprecedented media coverage of our business'. Despite the nature of that coverage, the number of users had increased by more than half to over 50 million. He also insisted that the company no longer used bots – indeed, that it was unnecessary since the new sign-ups were almost equal numbers of men and women. 'Our monthly new member account additions have not been verified by a third party,' he conceded, 'but we stand behind them.'[113] The company later went one step further and retained the accounting firm EY for what must have been one of its more unusual audits. In addition to verifying new user registrations, it confirmed that the 'Bot programs' had been decommissioned.[114]

The lawsuits and the FTC investigation against Ashley Madison were eventually settled.[115] The FTC imposed a $1.6m fine and a requirement for a comprehensive data security programme. For the class action, a final amount of $11.2m was put into a dedicated account. Because many users had signed up using fake email addresses, targeted banner advertisements were purchased to expand the class of potential claimants. Forty-two unnamed plaintiffs sought leave to take part in the action and settlement under pseudonyms – 'to reduce the risk of potentially catastrophic personal and professional consequences that could befall them and their families' should they be publicly identified. The court was sympathetic but denied this request, holding that their concerns about embarrassment did not outweigh the public's interest in transparency.[116]

[113] Richard Morgan, 'Ashley Madison Is Back – and Claims Surprising User Numbers', *New York Post* (21 May 2017).

[114] Ruby Life, Inc: Report on Customer Statistics for the Calendar Year 2017 (Ernst & Young LLP, 2018) 1.

[115] 'Operators of AshleyMadison.com Settle FTC, State Data Breach Charges' (2017) 34(3) Computer and Internet Lawyer 27.

[116] *In re Ashley Madison Customer Data Security Breach Litigation* MDL No 2669 (Eastern District of Missouri, Eastern Division, 6 April 2016).