13

Machine Learning, Cognitive Sovereignty and Data Protection Rights with Respect to Automated Decisions

Lee A. Bygrave

13.1 INTRODUCTION

In Shoshana Zuboff's recent treatise on 'surveillance capitalism', there is an evocative passage where she envisions a future hive of harmonious human–machine integration. In this future, she writes,

> [w]e will all be safe as each organism hums in harmony with every other organism, less a society than a population that ebbs and flows in perfect frictionless confluence, shaped by the means of behavioral modification that elude our awareness and thus can neither be mourned nor resisted.[1]

With a large dose of irony, Zuboff terms this future a 'utopia of certainty'. We are fortunately still some distance from it. However, as she and others document, an array of powerful economic, political and ideological forces propel us in its direction, and multiple technological-organisational processes provide 'the writing on the wall'.

Chief among such processes is the ever more pervasive automation of decisional systems governing human behaviour. We increasingly task computers and their program code to make or shape decisions that have direct and often significant effects on our well-being. Common examples are the automated ranking and selection of job applicants ('e-recruiting'),[2] government allocation of welfare payouts,[3] clinical decision-support systems,[4] prediction of the likelihood of recidivism in criminal justice contexts,[5] and curation of the newsfeeds of online social media

[1] S. Zuboff, *The Age of Surveillance Capitalism* (London: Profile Books, 2019), pp. 410–411.

[2] See further, e.g., E. Faliagka, A. Tsakalidis and G. Tzimas, 'An Integrated e-Recruitment System for Automated Personality Mining and Applicant Ranking' (2012) 22, no. 5 *Internet Research* 551–568. https://doi.org/10.1108/10662241211271545.

[3] See further, e.g., UN Special Rapporteur (2019). Report of the United Nations Special Rapporteur on Extreme Poverty and Human Rights, UN Doc.A/74/493 (11 October 2019) and references cited therein.

[4] See further, e.g., R. T. Sutton et al., 'An Overview of Clinical Decision Support Systems: Benefits, Risks, and Strategies for Success' (2020) 3 *NPJ Digital Medicine* 17. https://doi.org/10.1038/s41746–020-0221-y.

[5] A commonly discussed example being the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) software tool applied in the United States for guiding bail decisions: see further, e.g., J. Dressel and H. Farid, 'The Accuracy, Fairness, and Limits of Predicting Recidivism' (2018) 4 no. 1 *Science Advances* eaao5580. https://doi.org/10.1126/sciadv.aao5580. More generally, see W. L. Perry, B. McInnis, C. C. Price, S. C. Smith and J. S. Hollywood, *Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations* (Washington, DC: Rand Corporation, 2013).

platforms.[6] As the last-listed example highlights, automated decision making is applied not just in pursuit of the distribution of benefits, the meting out of penalties or the prediction of risk; it also shapes how we perceive ourselves and our environment.

At the same time, automated decision making is becoming 'smarter' and able to be gainfully employed in a growing number of contexts that require account to be taken of complex congeries of factors. This is because, in large part, of the application of sophisticated forms of artificial intelligence (AI) based on machine learning (ML) and 'Big Data' (BD). Although partly shrouded in hype and burdened by definitional disagreement,[7] these tools constitute key facilitators of societal change and look to remain so in coming years. In combination, they essentially involve the capacity to model some aspect of the world through data analysis, draw inferences from the model(s) in order to predict or anticipate possible events or behaviour, and improve their performance.

This chapter describes the basic mechanics of automated decisional systems, with a particular focus on those that employ ML and BD. It charts anxieties over their impact on human well-being, placing special emphasis on the challenges they pose for our interest in understanding our environs and ourselves, and for our moral and legal interest in doing so. This is precisely one of the interests that Zuboff's passage intimates as lost in her envisioned dystopia. Borrowing from terminology first suggested by Ulrich Beck,[8] I term the interest 'cognitive sovereignty'. Focus on this interest, I argue, fills a blind spot in scholarship and policy discourse on ML-enhanced decisional systems. More importantly, the interest is vital for grounding claims for greater explicability of machine processes.

Thereafter, the chapter casts a critical light on the newly revamped legal framework for data protection in the European Union (EU), particularly the General Data Protection Regulation (GDPR).[9] The chapter's primary focus is on those provisions of the GDPR that are aimed squarely at subjecting ML-enhanced and other automated decisional systems to greater human control. In addition to highlighting major points of controversy and uncertainty afflicting these provisions, the chapter assesses the degree to which they are able to fulfill their aim. I contend that they have considerable potential to ameliorate the potentially nefarious effects of automated decisional systems with a view to safeguarding cognitive sovereignty and related interests. However, I also argue that these provisions, on their own, constitute an overly narrow and ambiguous framework for tackling these effects, and that other more generally framed norms, both within and beyond the GDPR, may provide a stronger framework in this respect.

[6] See further, e.g., A. Chung, 'News feeds, old content: A brief history of algorithmically curated feeds on Facebook and Twitter' (2019). Available at https://medium.com/@annawchung/news-feeds-old-content-a-brief-history-of-algorithmically-curated-feeds-on-facebook-and-twitter-85b5e5d8e30a.

[7] See, e.g., M. Hildebrandt, 'The Artificial Intelligence of European Union Law' (2019) 21 *German Law Journal* 74–79. https://doi.org/10.1017/glj.2019.99, p. 74 (observing that 'AI' is 'a rather vague notion upon which no agreement exists, neither amongst experts nor amongst those affected by its supposedly disruptive character'); R. Kitchin and G. McArdle, 'What Makes Big Data, Big Data? Exploring the Ontological Characteristics of 26 Datasets' (2016) 3, no. 1 *Big Data and Society* 1–10. https://doi.org/10.1177/2053951716631130, p. 2 (noting that 'Big Data' as a term 'is being used as a catch-all label for a wide selection of data', with the result that it 'is treated like an amorphous entity that lacks conceptual clarity').

[8] U. Beck, *Risk Society: Towards a New Modernity* (London: Sage Publications, 1992). Originally published as *Risikogesellschaft. Auf den Weg in eine andere Moderne* (Frankfurt am Main: Suhrkamp Verlag, 1986), pp. 53–54.

[9] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] O.J. L 119/1.

13.2 MACHINE LEARNING: MECHANICS, CONSEQUENCES AND WORRIES

Old AI efforts[10] focused predominantly on creating so-called rule-based expert systems. These systems apply rigid, logical (e.g., 'if-then') rules that attempt to emulate human expertise in relatively narrow domains but with little if any inherent capacity to adapt or improve their performance.[11] In the latter respect, ML is a 'game changer'. Unlike static, rule-based expert systems, ML-enhanced AI involves programming computers to optimise performance (based on defined models)[12] using example/sample data, and where the basic task involves drawing inferences from the samples.[13] Deployment of an ML system is preceded first by a training phase (whereby the model in question is tested through application to a set of training data) and then a validation phase (whereby the model's classificatory parameters are refined through application to a second dataset). The learning dimension essentially inheres in the ability of the applied algorithms (i.e., the sequences of instructions in computer code that are used to transform input to output) to alter their output based on what they 'experience' in the form of feedback on their own input. This learning process takes various forms, differentiated in terms of the degree to which the learning is supervised,[14] and in terms of the inferential architecture (such as decision trees or neural networks)[15] employed.

A commonplace assumption is that most contemporary, automated decisional systems involve ML. In fact, the degree of ML uptake in this respect is modest. ML needs discretionary or logical 'space' in which to develop, and will thus be shut out of decisional systems where there is no such facility. Decisional systems that encode and apply static, non-discretionary legislative requirements are by their very nature incapable of utilising ML techniques, at least as a central part of their operations. This is the case for numerous automated decisional systems within government administration, such as those for calculating and paying social welfare benefits.[16]

[10] I use the term 'AI' in a generic sense to denote computer performance of tasks that ordinarily require human intelligence when carried out by people. This is in line with Kurzweil's influential conception of AI: see R. Kurzweil, *The Age of Intelligent Machines* (Cambridge, MA: MIT Press, 1990), p. 14 (defining AI as 'the art of creating machines that perform functions that require intelligence when performed by people'). However, as noted previously, the definition of AI is highly contested with multiple definitions proposed over the last three decades. For a concise overview, see S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (4th ed.) (Englewood Cliffs, NJ: Prentice Hall, 2020), ch. 1.

[11] R. Susskind, 'Artificial Intelligence and the Law Revisited' in: D. W. Schartum, L. A. Bygrave and G. G. Berge Bekken (eds.) *Jon Bing: En hyllest / Jon Bing: A tribute* (Oslo: Gyldendal, 2014), p. 197.

[12] In this context, a 'model' is a 'structure and corresponding interpretation that summarizes or partially summarizes a set of data, for description or prediction': R. Kohavi and F. Provost, 'Glossary of Terms' (1998) 30 *Machine Learning* p. 273.

[13] E. Alpaydin, *Introduction to Machine Learning* (Cambridge, MA: MIT Press, 2014); J. Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms (2016) 3, no. 1 *Big Data & Society* 1–12. https://doi.org/10.1177/2053951715622512.

[14] In this context, 'supervised' learning denotes 'techniques used to learn the relationship between independent attributes and a designated dependent attribute (the label)', whereas 'unsupervised' learning refers to techniques where no pre-specified dependent attribute is employed: Kohavi and Provost (note 12), p. 274.

[15] A decision tree in this context is basically a hierarchically structured, predictive classification model that maps observations about a particular unit of analysis to arrive at conclusions about its character. In contrast, a neural network (or, more accurately, artificial neural network) is a more complex, multi-layered architecture loosely modelled on the learning mechanisms of the human brain. Such a network consists of connected nodes ('neurons') that receive and transmit data between themselves in accordance with 'weight' thresholds for each connection, and where the network 'learns' by updating these weight thresholds in light of feedback on its output. See further, e.g., Russell and Norvig (note 10), ch. 21.

[16] For a systematic account of the various data-processing methods utilised in automated administrative decision making, see D. W. Schartum, 'From Facts to Decision Data: About the Factual Basis of Automated Individual Decisions' (2018) 50 *Scandinavian Studies in Law* 379–400.

Even outside this context, many of the latest 'hi-tech' endeavours do not involve ML: the case, for instance, with most of the automated 'track-and-trace' systems put in place by various European countries to help combat the COVID-19 pandemic.[17] Nonetheless, multiple opportunities for exploiting ML in decisional processes exist also in the field of government administration. ML may be utilised either at the front- or back-end of government decisional processes: for example, to make sense of aggregated datasets that constitute a legally relevant factual basis for decisions as to where to direct efforts at controlling misuse of government resources. I return to these sorts of opportunities further on in this chapter.

Improvement of ML typically leverages off access to large datasets that commonly go under the name 'Big Data'. Applying the conceptual framework advanced by Kitchin and McArdle, the key defining features of BD are *velocity* (BD are created quickly and in real time, and further processed quickly) and *exhaustivity* (BD capture entire systems – i.e., n = all – rather than capturing samples).[18] The marriage of ML and BD is often termed 'Big Data Analytics' (BDA). This sort of analytics privileges identifying correlations rather than causation. Much of the interest around BDA is rooted in its ability to find correlations in huge, relatively unstructured datasets that human cognition is ordinarily unable to discern. Summing up its basic mission, Cukier and Mayer-Schönberger state, 'Big Data helps answer what, not why, and often that's good enough.'[19] Nonetheless, we must remember that what is found is a statistical correlation (or, at best, covariation: i.e., a measure of the correlation's strength) based on a set of assumptions. Thus, ML-based predictions or classifications are essentially 'educated guesses or bets, based on large amounts of data'.[20] The assumptions they build on tend to express forms of bias (i.e., preferences for particular outcomes) and are accordingly far from value-neutral.[21] Moreover, as signaled by the well-known computer science maxim 'garbage in, garbage out', the utility of the findings depends on the reliability and validity of the sample data and the degree to which they – along with other applied metrics – are adequate proxies for the values they are supposed to represent.

Although intrinsically an epistemic enterprise, ML and BDA are involved not simply in generating or acting on knowledge. They function also as regulatory tools facilitating various forms of social ordering. An increasingly used generic term for this functionality is 'algorithmic regulation'. On its face, the term could denote regulation of algorithms but is supposed to denote regulation *by* algorithms. In what is becoming a standard definition, Karen Yeung describes algorithmic regulation as

> decision-making systems that regulate a domain of activity in order to manage risk or alter behaviour through continual *computational* generation of knowledge from data emitted and

[17] Algorithm Watch and Bertelsmannstiftung, *ADM Systems in the COVID-19 Pandemic: A European Perspective* (2020). Available at https://algorithmwatch.org/wp-content/uploads/2020/08/ADM-systems-in-the-Covid-19-pandemic-Report-by-AW-BSt-Sept-2020.pdf.

[18] Kitchin and McArdle (note 7).

[19] K. Cukier and V. Mayer-Schönberger. 'The Rise of Big Data: How It's Changing the Way We Think about the World' (2013) 92, no. 1 *Foreign Affairs* 29; V. Mayer-Schönberger, and K. Cukier, '*Big Data: A Revolution That Will Transform How We Live, Work, and Think* (Boston/New York: Houghton Mifflin Harcourt, 2013), ch. 4.

[20] T. Scantamburlo, A. Charlesworth and N. Cristianini, 'Machine Decisions and Human Consequences' in: K. Yeung and M. Lodge (eds.) *Algorithmic Regulation* (Oxford: Oxford University Press, 2019), pp. 49–81, at 57.

[21] J. E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford: Oxford University Press, 2019), p. 249; Scantamburlo, Charlesworth and Cristianini (note 20); G. Sartor and F. Lagioia, *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence* (Brussels: European Parliamentary Research Service, Scientific Foresight Unit (STOA) PE 641.530, 2020). Available at www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf, p.12.

> directly collected (in real time on a continuous basis) from numerous dynamic components pertaining to the regulated environment in order to identify and, if necessary, automatically refine (or prompt refinement of) the system's operations to attain a pre-specified goal.[22]

Algorithmic regulation may operate reactively but is increasingly used pre-emptively to predict and minimise unwanted behaviour or outcomes, or, concomitantly, to encourage desired (but not necessarily societally desirable) behaviour/outcomes. Traces of it are manifest across numerous contexts, including contracting,[23] medical care,[24] gambling,[25] curation of online content[26] and trading in financial markets.[27]

The decisional systems at the heart of algorithmic regulation, along with automated decision making more generally, are sometimes described as 'autonomous'. This description gives the misleading impression that they are impervious to human control. To be sure, many such systems have an inherent logic that can make their trajectories of use difficult for humans to understand and regulate (a difficulty to which I return later in the chapter). Yet, this logic – as in the case of other technologies – embodies the values of its human creators, and these values lay constraints on the technologies' use.[28] Mireille Hildebrandt aptly expresses the human-dependence of machine intelligence (at least in its present form) as follows,

> Machines cannot do anything but execute programs developed by humans, even if those programs enable the machine to reconfigure its program in view of specified machine-readable tasks, and even if humans may develop programs that build new programs.[29]

However, the long-term possibility of machine intelligence developing into less human-dependent forms should not be discarded. Numerous factors might play a role in this regard, although their effects are hard to predict. One such factor is the ongoing development of quantum computing and neuromorphic computing, each of which promises huge leaps in computers' efficiency, flexibility and processing power.[30] Another is the combination of ML systems in complex networks that could produce an intelligence far greater than the sum of their components.

ML-enhanced decision making is typically sold on the promise of significant gains in decisional speed, scalability, consistency, predictability and accuracy, and this may in turn have

[22] K. Yeung, 'Algorithmic Regulation: A Critical Interrogation' (2018) 12, no. 4 *Regulation & Governance* 505–523, at 507. https://doi.org/10.1111/rego.12158.

[23] See, e.g., H. R. Varian, 'Computer Mediated Transactions' (2010) 100, no. 2 *American Economic Review* 1–10 (describing 'computer-mediated' contracts that are automatically enforced or terminated on the basis of continuous automated monitoring of the degree to which they are complied with).

[24] See, e.g., Z. Obermeyer and E. J. Emanuel, 'Predicting the Future: Big Data, Machine Learning, and Clinical Medicine' (2016) 375, no. 13 *The New England Journal of Medicine* 1216–1219. https://doi.org/10.1056/NEJMp1606181; A. Avati et al., 'Improving Palliative Care with Deep Learning' (2018) 18, suppl. 4 *BMC Medical Informatics and Decision Making* 55–64. https://doi.org/10.1186/s12911–018-0677-8.

[25] See, e.g., N. Dow Schull, *Addiction by Design* (Princeton, NJ: Princeton University Press, 2012).

[26] See, e.g., Chung (note 6).

[27] See, e.g., T. Myklebust 'Fairness and Integrity in High-Frequency Markets: A Critical Assessment of the European Regulatory Approach' (2020) 31, no. 1 *European Business Law Review* 33–76.

[28] See further, e.g., the classic analysis in M. Akrich, 'The De-scription of Technical Objects' in: W. Bijker and J. Law (eds.) *Shaping Technology / Building Society: Studies in SocioTechnical Change* (Cambridge, MA: MIT Press, 1992), pp. 205–224 (describing the 'inscription' processes by which engineers and designers embed their visions of future users in the technical objects they create).

[29] Hildebrandt (note 7).

[30] E. Knill, 'Quantum Computing' (2010) 463 *Nature* 441–443. https://doi.org/10.1038/463441a; P. Stone et al., Artificial intelligence and life in 2030. *Report of the 2015 Study Panel for the One Hundred Year Study on Artificial Intelligence*, 2016. Available at https://ai100.stanford.edu/sites/g/files/sbiybj9861/f/ai_100_report_0831fnl.pdf.

positive macro-level effects, such as improved organisational efficiency, profit margins, service levels and productivity.[31] These benefits may accrue not just to the entities employing automated decisional systems but also to those who are the decision subjects and to broader society. In practice, though, shortfalls in delivering such benefits occur frequently, not least in the provision of government services.[32] This is the result partly of persistent difficulties in insulating machine processes from human error, inefficiency and other 'foibles' in the organisational framework for their design and application.[33]

Even when it delivers handsomely, ML-enhanced decision making and automated decision making more generally may pose significant threats, both for the decision subjects and for the health of wider society. In large part, the threats arise from the potential for such decision making to be a vehicle for unwarranted bias, discrimination, gaming, addiction or manipulation.[34] The gravity of these threats has grown in line with the increasing definitional power of inferencing algorithms. In the witty phrasing of George Dyson, 'Facebook defines who we are, Amazon defines what we want, and Google defines what we think.'[35] Although the witticism contains an element of hyperbole, its gist is valid. This has worrisome implications for humans' self-perception and 'Weltanschaung'. It also has alarming political implications, not least for the quality of democratic electoral processes: consider, for instance, the recent scandals involving covert political micro-targeting of Facebook users through use of 'dark ads'.[36]

Another disturbing exemplification of inferencing algorithms' definitional power is in the machinations of the 'digital welfare state' where there is ever greater reliance on computer programs to identify and prosecute persons who are supposedly 'rorting the system', and where the putatively guilty must go to great lengths to prove their innocence even when solid evidence of error exists.[37] This is particularly problematic given that the decision subjects are usually

[31] See, e.g., T. O'Reilly, 'Open data and algorithmic regulation' in: B. Goldstein and L. Dyson (eds.), *Beyond Transparency: Open Data and the Future of Civic Innovation* (San Francisco, CA: Code for America Press, 2013), ch. 22. Available only online at: https://beyondtransparency.org/chapters/part-5/open-data-and-algorithmic-regulation/; European Commission, White Paper: On Artificial Intelligence - A European approach to excellence and trust. COM (2020) 65 final (2020).

[32] See, e.g., A. Griffiths, 'The Practical Challenges of Implementing Algorithmic Regulation for Public Services' in: K. Yeung and M. Lodge (eds.) *Algorithmic Regulation* (Oxford: Oxford University Press, 2019), ch. 7; M. Zalnieriute, L. Bennett Moses and G. Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82, no. 3 *Modern Law Review* 425–455. https://doi.org/10.1111/1468-2230.12412.; M. Zalnieriute et al., 'From Rule of Law to Statute Drafting: Legal Issues for Algorithms in Government Decision-Making' in: W. Barfield (ed.) *Cambridge Handbook on the Law of Algorithms* (Cambridge: Cambridge University Press, 2020), pp. 251–272.

[33] Griffiths (note 32).

[34] See further, e.g., I. Kerr and J. Earle, 'Prediction, Preemption, Presumption: How Big Data Threatens Big Picture' (2013) 66 *Stanford Law Review Online* no pagination. Available at www.stanfordlawreview.org/online/privacy-and-big-data-prediction-preemption-presumption/; K. Yeung, 'Hypernudge': Big Data as a Mode of Regulation by Design' (2017) 20(1), *Information, Communication & Society* 118–136. https://doi.org/10.1080/1369118X.2016.1186713; K. Yeung, 'Why Worry about Decision-Making by Machine?' in: K. Yeung and M. Lodge (eds.) *Algorithmic Regulation* (Oxford: Oxford University Press, 2019), ch. 2; UN Special Rapporteur, Report of the United Nations Special Rapporteur on Extreme Poverty and Human Rights, UN Doc.A/74/493 (2019).

[35] G. Dyson, *Turing's Cathedral: The Origins of the Digital Universe* (New York: Pantheon, 2012), p. 308.

[36] S. Vaidhyanathan, *Anti-Social Media: How Facebook Disconnects Us and Undermines Democracy* (New York: Oxford University Press, 2018); F. J. Zuiderveen Borgesius et al., 'Online Political Microtargeting: Promises and Threats for Democracy' (2018) 14, no. 1 *Utrecht Law Review* 82–96. https://doi.org/10.18352/ulr.420.

[37] See, e.g., the scandal surrounding the automated debt recovery ('robo-debt') scheme operated by the Australian Government Department of Human Services: M. Zalnieriute, L. Bennett Moses, and G. Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82, no. 3 *Modern Law Review* 425–455, at 436–437. https://doi.org/10.1111/1468-2230.12412.

among the least able to stand up for their rights.[38] Moreover, it is problematic for the foundational values of the welfare state. These values focus on the livelihood conditions of a state's population and typically establish an individualised, rights-based point of departure for provision of welfare.[39] Automated control processes of the sort outlined threaten to undermine that foundation. Additionally, they detract from 'rule of law' ideals.[40]

BDA-supported practices also carry risks well beyond the sphere of social welfare and civil liberties. In respect of market dynamics, for instance, 'price-bots' and 'computer cartels' may facilitate anti-competitive collusion between market participants,[41] while 'high frequency trading' may threaten the integrity and long-term societal value of financial markets.[42]

A pervasive and especially acute worry across most if not all contexts in which ML-enhanced decisional processes operate is the opacity of their mechanics and logic. The problem of opacity reflects not simply shortfalls in humans' computer programming skills but the fact that the decisional processes involved do not closely emulate the logic of human thought processes. Further, as the algorithms 'learn' in successive feedback loops, their decisional logic becomes less tethered to its point(s) of departure. These factors, combined with the sheer immensity (and often considerable heterogeneity) of the datasets involved, may result in a complexity of logic that defies ready human interpretability, even for experts,[43] thus making the decisional processes a type of 'black box'.[44] There is a paradox here given the basic epistemic mission of ML and BD. Richards and King term this a 'transparency paradox': 'Big data promises to use . . . data to make the world more transparent, but its collection is invisible, and its tools and techniques are opaque.'[45] This informational imbalance engenders a control potential resembling the classic panoptic dynamic described by Foucault.[46] Yet, as Yeung reminds us, this potential 'is both more potent and powerful than the kind of disciplinary control typically associated with pre-digital forms of surveillance which rely upon the coercive experience of living with the uncertainty of being seen'.[47]

Organisations' self-interest exacerbates the opacity problem. For many organisations, private or public, their algorithms are akin to a secret sauce they will go to great lengths to protect from disclosure.[48] Inasmuch as the algorithms qualify as trade secrets or intellectual property, laws

[38] UN Special Rapporteur (note 34), para. 63 ('Digital technologies are employed in the welfare state to surveil, target, harass and punish beneficiaries, especially the poorest and most vulnerable among them').

[39] I. Ikdahl and V. Blaker Strand, *Rettigheter i Velferdsstaten. Begreper, Trender, Teorier* (Oslo: Gyldendal, 2016).

[40] Zalnieriute, Bennett Moses and Williams (note 32); M. Finck, 'Automated Decision-Making and Transparency in Administrative Law' in: P. Cane (ed.) *The Oxford Handbook on Comparative Administrative Law* (Oxford: Oxford University Press, 2021), pp. 657–676.

[41] A. Ezrachi and M. Stucke, *Virtual Competition: The Promise and Perils of the Algorithm-Driven Economy* (Boston, MA: Harvard University Press, 2016); S. Schechner, 'Why do gas station prices constantly change? Blame the algorithm' (2017) *Wall Street Journal*, 8 May 2017. Available at www.wsj.com/articles/why-do-gas-station-prices-constantlychange-blame-the-algorithm-1494262674.

[42] Myklebust (note 27); W. Mattli, *Darkness by Design: The Hidden Power in Global Capital Markets* (Princeton, NJ: Princeton University Press, 2019).

[43] See further, e.g., Burrell (note 13), p. 2 (elaborating on what Burrell terms the 'mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation').

[44] D. Castelvecchi, 'Can We Open the Black Box of AI?' (2016) 538 *Nature* 20–23.

[45] N. M. Richards and J. H. King, 'Three Paradoxes of Big Data' (2013) 66 *Stanford Law Review Online* no pagination. Available at www.stanfordlawreview.org/online/privacy-and-big-data-three-paradoxes-of-big-data/.

[46] M. Foucault, *Discipline and Punish: The Birth of the Prison* (Harmondsworth: Penguin, 1977), pp. 195–228.

[47] Yeung (note 34), p. 130.

[48] For exemplification of private corporations' preference to keep their algorithms secret, see F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Cambridge, MA: Harvard University Press, 2015), ch. 3. For exemplification of government agencies' non-disclosure practices, see the refusal of the Dutch state to reveal key details of the logic and mechanics of its 'SyRI' fraud-detection system, referenced in Section 13.4.

safeguarding such secrets or property may also stymy efforts to open up the algorithms' logic to public scrutiny.[49] Indeed, the role played by such laws in this regard may arguably be seen as part of a long-term endeavour aimed at insulating or 'encasing' tools of informational capitalism from democratic accountability.[50]

We must be careful, though, not to paint boxes blacker than they really are nor to treat all boxes as equally or always black. The degree to which ML-based decisional processes are opaque varies according to the computational learning model they employ. A model based on a fairly simple decision tree is more amenable to explanation than one based on a non-linear, non-deductive, neural network for 'deep learning'. It is also worth keeping in mind that many computer scientists, mathematicians and others engaged in developing ML are increasingly making systematic efforts to develop methods to reduce or compensate for the blackness of, inter alia, neural networks. These efforts are commonly described using the nomenclature XAI ('eXplainable AI').[51] While they face an uphill struggle – not least owing to the aforementioned desire of many organisations to keep their 'sauces' secret – they show that the black-box problem is being tackled seriously by some of those who are partly responsible for it. Additionally, the encasement ability of laws protecting trade secrets and intellectual property with respect to computer algorithms is arguably not as strong as sometimes assumed, at least in Europe – a point elaborated in Section 13.4. Finally, we must not forget that decisional opacity is far from unique to ML. Decisional black boxes existed long before the age of BDA (recall, for example, the arcane processes of credit scoring in the United States before the 1970s),[52] and they exist today independently of BDA (recall, for instance, the opaque cronyism governing decisions about which families are offered places in public childcare facilities in France).[53] Ultimately, as Malcolm Langford points out, '[d]iscussions of automation and digitalization should be guided by a logic of minimizing danger, regardless of whether its origin is machine or human'.[54]

Nonetheless, the opacity of ML-based decisional systems is generally recognised as a real challenge for efforts to make them understandable and publicly accountable. There is also voluminous, insightful scholarship setting out various rationales for such efforts. Yet, despite its richness, this scholarship has fallen short in elucidating an important human interest that the opaque intelligence of ML-based decisional systems threatens: the interest in cognitive sovereignty. The interest is foundational to the normative justification for requiring explicability of machine processes. Section 13.3 elaborates on its substance.

13.3 COGNITIVE SOVEREIGNTY: FROM SHADOWS TO LIMELIGHT

The notion of cognitive sovereignty as used herein denotes a human being's ability and entitlement to comprehend with a reasonable degree of accuracy their environs and their place

[49] Finck (note 40) , 667.

[50] A. Kapczynski, 'The Law of Informational Capitalism' (2020) 129 *Yale Law Journal* 1460–1515, pp. 1508ff; J. E. Cohen, *Between Truth and Power: The Legal Constructions of Informational Capitalism* (Oxford: Oxford University Press, 2019), pp. 62–63.

[51] See, e.g., Defense Advanced Research Projects Agency, Explainable Artificial Intelligence (XAI). DARPA-BAA-16-53, 2016. Available at www.darpa.mil/attachments/DARPA-BAA-16-53.pdf; T. Miller, 'Explanation in Artificial Intelligence: Insights from the Social Sciences' (2017) 267, *Artificial Intelligence* 1–38. https://doi.org/10.1016/j.artint.2018.07.007.

[52] Pasquale (note 48), p. 22.

[53] A. Léchenet, 'In French daycare, algorithms attempt to tackle cronyism' (2020) *AlgorithmWatch*, 18 September 2020. Available at https://algorithmwatch.org/en/story/algorithms-to-fight-cronyism-in-french-daycare/.

[54] M. Langford, 'Taming the Digital Leviathan: Automated Decision-Making and International Human Rights' (2020) 114 *American Journal of International Law Unbound* 141–146, p. 145.

therein, particularly the implications these hold for their exercise of choice. The notion derives in part from sociological scholarship in the 1980s on the increasing degree to which consciousness of 'risk' (i.e., the possibility of human action triggering events with detrimental social consequences) pervades human interaction and self-perception in contemporary society. Central in this scholarship is the work of Ulrich Beck, who argued that a major distinguishing feature of modernity is the weighing down of human behaviour by a growing awareness of threat, vulnerability and unpredictability. In his view, we feel less able to discern what is dangerous and what is safe for ourselves, and, accordingly, we are ever more reliant on 'external knowledge producers' to gauge the degree to which we are endangered.[55] Beck summed up this development in terms of a gradual loss of 'cognitive sovereignty' ('Wissenssouveränität') over the parameters of our actions.[56]

Beck did not describe in detail what this sort of sovereignty entails. Nor did he flag its fate specifically in the face of ML – understandably, given the time at which he was writing. He made clear, though, that such sovereignty involves people being able to rely predominantly on themselves – more specifically, their 'own cognitive means and potential experiences' – to ascertain the extent of their endangerment. It is, accordingly, a state in which people are not 'incompetent ['unzuständig'] in matters of their own affliction'.[57]

Beck's notion of 'cognitive sovereignty' seems to have received little attention from other scholars in the years since its conception, apart from limited recognition of its importance with respect to data protection. Drawing partly on Beck's work, a treatise on data protection law written two decades ago noted deficits in our 'cognitive sovereignty' owing to 'data on ourselves … being handled by many persons and organisations of which we know little or nothing'.[58]

The treatise went on to advance the claim that an important remit of data protection law is to reduce these deficits and thereby shore up public trust in the way organisations process personal data.[59] It further noted manifestations of this remit in the core principles of such law, particularly the so-called purpose limitation principle requiring personal data to be collected for specified legitimate purposes and not used in ways that are incompatible with those purposes.[60] However, the treatise neither fleshed out the notion of cognitive sovereignty in any significant way nor raised it in relation to ML specifically.

It is now high time that we take the notion down from the shelf, dust it off, rejig it and give it work to do. In an age in which decisional processes that impact human well-being are increasingly steered by opaque logic, cognitive sovereignty must take centre stage as a key interest at stake. A thorough analysis of the interest is beyond the scope of this chapter; what follows is simply a preliminary account of its basic character. While this account builds on Beck's work, it goes beyond and in certain respects differs from his depiction of cognitive sovereignty.

To begin with, cognitive sovereignty is not to be confused with 'cognitive liberty', which has received a relatively large amount of scholarly attention. Cognitive liberty has been defined as 'the right to choose one's own cognitive processes, to select how one will think, to recognise that

[55] Beck (note 8), pp. 53–54.
[56] Ibid., p. 53.
[57] Ibid.
[58] L. A. Bygrave, *Data Protection Law: Approaching Its Rationale, Logic and Limits* (The Hague: Kluwer Law International, 2002), p. 111.
[59] Ibid.
[60] Ibid., p. 337.

the right to control thinking processes is the right of each individual person'.[61] Thus, cognitive liberty is essentially about the autonomy of one's mind or 'forum internum'.[62] In contrast, cognitive sovereignty concerns one's ability to comprehend one's *external* environment and one's place in it.[63] Yet, despite the distinction drawn between them, it will be readily apparent that cognitive liberty and cognitive sovereignty have a symbiotic relationship: cognitive liberty is a means of safeguarding cognitive sovereignty, just as the latter helps to maintain the former. Much the same can be said of the relationship between cognitive sovereignty and particular other recently conceived forms of sovereignty, notably 'technological sovereignty' and 'algorithmic sovereignty' as defined by, respectively, the Democracy in Europe Movement 2025 (DiEM25)[64] and Reviglio and Agosti.[65]

Cognitive sovereignty engages with the conditions for acquiring knowledge of the world (the 'cognitive' element) and presumes that each human has a moral and legal entitlement in being able to understand – more or less – their environs and how these impact on them (the 'sovereignty' element). The reference to 'more or less' underlines that the factual, moral and legal extents of cognitive sovereignty are inevitably matters of degree, not absolutes, and vary from context to context. Hence, cognitive sovereignty does not equate with omniscience, which belongs to the mythical realm of gods. This does not undermine the value of cognitive sovereignty as a moral or legal claim.[66]

As indicated previously, cognitive sovereignty is more than just a capacity and interest; 'sovereignty' denotes entitlement. It thereby signals that humans *deserve* the capacity to comprehend their environs and their place therein. The basis for entitlement is humans' status as sentient beings with innate dignity. An intrinsic aspect of this status is self-awareness – that is, our sense of identity. Safeguarding our cognitive sovereignty is a *sine qua non* for our perception of self. Through knowledge of the world and our place in it, we gain knowledge of who we are relative to others. Notwithstanding its dignitarian foundation, a more pragmatic basis for cognitive sovereignty exists as well: without a significant degree of such sovereignty, we would be hard-pressed to survive in the face of the numerous dangers to our biological lives.

On its face, cognitive sovereignty may be fitted snugly within the pantheon of classical liberal ideals that build on a conception of the human being as essentially autonomous, self-regarding

[61] See A. Weil, *The Natural Mind: An Investigation of Drugs and the Higher Consciousness* (Boston, MA: Houghton Mifflin Company, 1998), p. 140. The definition is cited and adopted in legal scholarship: see, e.g., C. Walsh, 'Drugs and Human Rights: Private Palliatives, Sacramental Freedoms and Cognitive Liberty' (2020) 14, no. 3 *The International Journal of Human Rights* 425–441, p. 433. https://doi.org/10.1080/13642980802704270.

[62] J. C. Bublitz and R. Merkel, 'Crimes against Minds: On Mental Manipulations, Harms and a Human Right to Mental Self-Determination' (2014) 8 *Criminal Law and Philosophy* 51–77, p. 64. https://doi.org/10.1007/s11572–012-9172-y

[63] See, however, K. Yeung, *A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility within a Human Rights Framework* (Strasbourg: Council of Europe, 2018), p. 80 (employing 'cognitive sovereignty' in much the same way as the notion of 'cognitive liberty' as described in this section, and, concomitantly, drawing on the work of Bublitz rather than the work of Beck or Bygrave).

[64] See Democracy in Europe Movement 2025 (DiEM25) (2019). Progressive agenda for Europe – Technological Sovereignty: Democratising Technology and Innovation (Green Paper No. 3). Available at https://diem25.org/wp-content/uploads/2019/03/Technological-Sovereignty-Green-Paper-No-3.pdf, p. 4 (defining 'technological sovereignty' as 'the right and capacity by citizens and democratic institutions to make self-determined choices on technologies and innovation').

[65] See U. Reviglio and C. Agosti, 'Thinking Outside the Black-Box: The Case for 'Algorithmic Sovereignty' in Social Media' (2020) 6, no. 2 *Social Media + Society* 1–12, p. 5. https://doi.org/10.1177/2056305120915613 (defining 'algorithmic sovereignty' as 'the moral right of a person to be the exclusive controller of one's own algorithmic life and, more generally, the right and capacity by citizens as well as democratic institutions to make self-determined choices on personalization algorithms and related design choices').

[66] See too the analogous line taken by Bublitz and Merkel (note 62), 65–66 in respect of mental self-determination.

and worthy of respect as an end rather than a means. This sort of conception has been extensively criticised for embracing a 'hyper-individualist' idealisation of the human self as 'fundamentally separated or able to easily become separated from all connections with intimate others, surrounding culture, and other identity-constituting elements of the social environment'.[67] Such an idealisation has little correspondence to reality and fails to recognise adequately the myriad ways in which a person's environment, over which they have limited control, shapes the exercise of their choice.[68] Thus, recent years have seen the growth of various 'relational', 'socio-relational' or 'communitarian' conceptualisations of autonomy.[69] These view the human individual as inextricably embedded in, and formed by, a web of social relations, and accordingly frame autonomy by reference to these. The account of cognitive sovereignty given here is compatible with this sort of conceptual framework. In other words, it recognises that the factual, moral and legal extents of a person's cognitive sovereignty are not only relative but contingent on the web of social relations in which the person is enveloped.

In this regard, the account differs from Beck's work inasmuch as he seemed to indicate (as noted earlier) that cognitive sovereignty only pertains when a person is able to comprehend their situation relying basically on their 'own cognitive means and potential experiences', and not those of 'external knowledge producers'. If this is Beck's view,[70] it pitches cognitive sovereignty well above what is practicably possible. In reality, a person's cognitive sovereignty depends in large part on external knowledge producers; the decisive factor is the extent to which the person is able to access and utilise that knowledge (assuming it is reliable). That factor depends, in turn, on the interpersonal and other social connections of the person. In other words, 'relational' elements are an important part of the background requirements for the sovereignty interest at stake.

Another characteristic setting apart the interest in cognitive sovereignty as described here from classical liberal ideals is that it is not just an interest that a person has *qua* individual; the interest has an aggregate dimension as well such that it attaches to collective entities, both those that are internally organised and those that are not. Thus, it speaks to the ability of, say, a particular profession, class, neighbourhood or ethnic group to understand the parameters of decisional systems that may affect their collective well-being. The aggregate dimension of the interest is not unique; for instance, most of the interests that data protection law seeks to safeguard can be shared by collective entities.[71] However, these interests' dignitarian rationale becomes weaker (but does not necessarily disappear) as the moral and legal connection between a collective entity and its human constituency weakens. That connection is typically weak in the case of large companies operating with a strong 'corporate veil', but considerably less so with families, households and many small enterprises. Nonetheless, for any collective entities at relatively high risk of being subject to undue discrimination or other unwarranted interference, the need to understand the parameters of their vulnerability looms large even if it is not always closely or prominently tied to dignitarian factors.

[67] J. Christman, 'Relational Autonomy, Liberal Individualism, and the Social Constitution of Selves' (2004) 117, no. 1/2 *Philosophical Studies* 143–164, p. 147. https://doi.org/10.1023/B:PHIL.0000014532.56866.5c.

[68] See Yeung (note 34).

[69] See further, e.g., J. Nedelsky, 'Reconceiving Autonomy: Sources, Thoughts and Possibilities' (1989) 1 *Yale Journal of Law and Feminism* 7–36; M. A. L. Oshana 'Personal Autonomy and Society' (1998) 29, no. 1 *Journal of Social Philosophy* 81–102. https://doi.org/10.1111/j.1467-9833.1998.tb00098.x; C. Mackenzie and N. Stoljar, 'Introduction: Autonomy Reconfigured' in: C. Mackenzie and N. Stoljar (eds.) *Relational Autonomy: Feminist Perspectives on Autonomy, Agency, and the Social Self* (New York: Oxford University Press, 2000), pp. 3–31.

[70] He is not crystal clear on this point: see Beck (note 8).

[71] See Bygrave (note 58), chs. 7, 12, 15.

The account of cognitive sovereignty set out herein is far from revolutionary either in terms of the interest it denotes or the justification for it. With respect to justification, Tal Zarsky, for example, has provided a rationale for decisional transparency that aligns closely to the dignitarian justification provided for cognitive sovereignty.[72] With respect to the interest concerned, Norwegian scholars, for instance, argued as far back as the early 1970s that a person has an interest in 'insight' ('innsyn') or 'awareness' ('opplysthet') concerning who processes data about themselves, what data are processed and the purpose(s) for the processing.[73] Additionally, there is, of course, a much longer heritage of discourse on the interlinked principles of natural justice, rule of law and due process which is concerned with the interest in ensuring that government bodies' decision making is transparent, foreseeable and not arbitrary.

So do we really need to employ the notion of 'cognitive sovereignty' as explicated in this section in discourse on processes that challenge people's ability to understand what is happening to them? Is it not just a cumbersome, pretentious expression for an obvious and simple interest that other more commonly used terms may cover? In my view, it fills a gap in the standard list of terms that scholars trot out whenever they attempt to describe the basic human interests threatened by AI, ML, BD and the like. References to 'autonomy', 'personal integrity', 'dignity' and 'privacy' dominate this scholarly discourse. While nebulous and pregnant with definitional variation, all of these terms are semantically poor proxies for cognitive sovereignty in the way it is defined herein. As noted, cognitive sovereignty is distinct from autonomy even if its realisation may depend on the latter. It is also obviously distinct from integrity and dignity inasmuch as these terms connote, respectively, a state of non-interference and the innate worth of humans along with their concomitant treatment as ends rather than means. As for privacy, this connotes a large range of overlapping conditions, claims and interests,[74] none of which are commensurate with cognitive sovereignty as defined herein.

Regarding the desired nomenclature, a variety of other, seemingly simpler terms, such as 'awareness' or 'insight', could do the work of 'cognitive sovereignty'. Yet, though cumbersome, the latter has a gravity and pondus that may be useful in advocacy for greater explicability of ML-enhanced (or other arcane) decisional processes. This is particularly so in light of the previously noted reluctance of many organisations employing such processes to disclose details of their logic and mechanics. Paul Freund remarked many years ago that adopting a 'large concept' such as 'privacy' could be useful 'in order to offset an equally large rhetorical counter-claim: freedom of inquiry, the right to know, liberty of the press . . . '.[75] The same dynamic may apply with respect to use of 'cognitive sovereignty' as a rhetorical device to challenge weighty counter-interests.

13.4 DATA PROTECTION RIGHTS PERTAINING SPECIFICALLY TO AUTOMATED DECISIONS

The account of cognitive sovereignty set out in Section 13.3 shows the interest as possessing an existential compass. Thus, the interest comes into play not simply in the context of humans'

[72] T. Zarsky (2013). Transparency in Data Mining: From Theory to Practice in: B. Custers, T. Calders, B. Schermer and T. Zarsky (eds.) *Discrimination and Privacy in the Information Society. Studies in Applied Philosophy, Epistemology and Rational Ethics*, vol. 3 (Berlin/Heidelberg: Springer, 2013), pp. 301–324, at 317 ('An individual has a right to learn the reasons for events which affect him. Such information empowers her, and she senses she is treated with respect, as a human being.')

[73] See Bygrave (note 58), p. 140 and references cited therein.

[74] See the comprehensive classification in B.-J. Koops et al., 'A Typology of Privacy' (2017) 38 *University of Pennsylvania Journal of International Law* 483–575.

[75] P. A. Freund, 'Privacy: One Concept or Many' in: J. R. Pennock and J. W. Chapman (eds.) *Privacy: Nomos XIII* (New York: Atherton Press, 1971), pp. 182–198.

ability to understand the parameters of the processing of data about themselves by other persons or entities – a context central to the field of data protection law – but across multiple dimensions of human activity, both at the individual, micro- or local level and the collective, macro- or systemic level. Nonetheless, it is clear that the interest is forcefully engaged by data-processing practices, particularly those connected with the decisional processes outlined earlier in the chapter, and that much of the data involved can be linked to identifiable, individual natural/physical persons, meaning that the data are 'personal' and accordingly fall within the ambit of data protection law.[76] This means, in turn, that data protection law is a significant determinant of cognitive sovereignty's prospects in the face of automated decision making directed at human beings.

One of the aims behind the recent overhaul of the EU's legislative framework for data protection was to create better safeguards for fundamental rights and freedoms – including, implicitly, those related to cognitive sovereignty – in light of current or emerging technological-organisational realities. ML and BDA as such did not figure prominently in the reform negotiations, but EU lawmakers appear to have regarded these processes as part of the challenges to be tackled by the GDPR,[77] although they ultimately subsumed them within the overlapping categories of automated decision making and profiling. Today, the GDPR is commonly portrayed as playing a key role in seeking to secure algorithmic accountability. The European Commission, for example, pitches adherence to the GDPR as a vital prerequisite for realising an 'ecosystem of trust' for AI in which the EU's values are respected.[78] As elaborated in this section, however, it is questionable whether EU lawmakers, when drafting the GDPR, fully appreciated all of the regulatory challenges thrown up by ML and BDA, not least the potential conundrum facing decision makers who are obliged to explain the logic of a ML-enhanced decisional process but struggle themselves to understand that logic.

Article 22 GDPR is central to the regulation of ML-enhanced decisional processes. This gives a person several rights with respect to automated decision making, chief of which is a qualified right 'not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her'

[76] See the definition of 'personal data' – a key boundary concept for the application of data protection law – in Article 4 (1) GDPR. For elaboration, see, e.g., L. A. Bygrave and L. Tosoni, 'Article 4(1): Personal Data' in: C. Kuner, L. A. Bygrave, C. Docksey and L. Drechsler (eds.), *The EU General Data Protection Regulation* (Oxford: Oxford University Press, 2020), pp. 103–115.

[77] For example, the European Commission's initial proposal for the GDPR took explicit account of the Council of Europe's 2010 Recommendation on data protection in the context of profiling: European Commission, 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) (COM (2012) 11 final), p. 9 (fn. 33). The preamble to the recommendation refers to, inter alia, 'calculation, comparison and statistical correlation software, with the aim of producing profiles' and 'automatic application of pre-established rules of inference'. The Explanatory Memorandum to the recommendation also contains passages dealing implicitly with BDA, such as the following: 'With statistical tools and algorithms, it thus becomes possible to identify connections between certain kinds of behaviour. Human common sense and logic play no part in establishing these correlations. It is purely the computing power and the sophistication of the algorithms that bring to light correlations often invisible to the naked eye or beyond human reason, albeit without explaining them' (Council of Europe, Recommendation Cm/Rec(2010)13 on the protection of individuals with regard to automatic processing of personal data in the context of profiling. Strasbourg: Council of Europe Publishing, p. 39 (para. 97)). Recital 71 GDPR also intimates concern about BDA ('the controller should use appropriate mathematical or statistical procedures for … profiling' in order to minimise 'risk of errors' and prevent 'discriminatory effects').

[78] European Commission, White Paper: On Artificial Intelligence – A European approach to excellence and trust. COM(2020) 65 final (19 February 2020), especially pp. 3, 9, 16, 19.

(Article 22(1)). Other recently enacted European data protection laws, such as the Council of Europe's modernised Convention on data protection ('Convention 108+')[79] and the EU's Law Enforcement Directive ('LED'),[80] place similar (but not uniformly commensurate) restrictions on automated decision making.

These restrictions' normative roots stretch across at least two earlier generations of data protection law,[81] and arguably even further back to rules on due process and 'natural justice' in administrative and criminal law. However, their direct antecedents – such as Article 15 of the former EU Data Protection Directive[82] – were limited to situations involving the application of personal profiles. Article 22 GDPR and its equivalents in Convention 108+ and the LED are not; they embrace decisions based on 'automated processing' of which 'profiling' is but one example.[83] Regardless, each generation of these norms is grounded in much the same sort of worry: first and foremost, fear for the future of human dignity in the face of machine determinism. More specifically, they are each rooted in a concern to ensure that humans are able to participate in, shape and retain responsibility for decisions that significantly impact other humans, and that humans accordingly maintain the primary role in 'constituting' themselves.[84] Anxieties over the risk of machine error and undue discrimination augment this concern.[85]

A need to uphold cognitive sovereignty lies implicit in the policy underpinnings for Article 22 but does not occupy a prominent place. It shines more clearly in other provisions of the GDPR that also deal specifically with fully automated decision making. These provisions are Articles 13(2)(f), 14(2)(g) and 15(1)(h) GDPR, all of which require (albeit in differing contexts) that data subjects be informed of 'the existence' of the decisional processes caught by Article 22 (1) and, 'at least in those cases', be provided with 'meaningful information about the logic

[79] Modernised Convention for the Protection of Individuals with regard to the Processing of Personal Data (consolidated text), adopted by the 128th Session of the Committee of Ministers, 17–18 May 2018. See particularly Article 9(1)(a).

[80] Directive (EU) 2016/680 on the protection of natural persons with regard to the processing of personal data by competent authorities for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, and on the free movement of such data, and repealing Council Framework Decision 2008/977/JHA [2016] O.J. L 119/89. See particularly Article 11.

[81] See further L. A. Bygrave, 'Minding the Machine v2.0: The EU General Data Protection Regulation and Automated Decision-Making' in: K. Yeung and M. Lodge (eds.) *Algorithmic Regulation* (Oxford: Oxford University Press, 2019), pp. 246–260, at 249, 252.

[82] 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 [1995] O.J. L 281/31 (repealed). See further L. A. Bygrave, 'Minding the Machine: Article 15 of the EC Data Protection Directive and Automated Profiling' (2001) 17, no. 1 *Computer Law & Security Review* 17–24. https://doi.org/10.1016/S0267–3649(01)00104-2.

[83] Some commentators argue otherwise – i.e. that profiling is a necessary component of decision making captured by Article 22(1) GDPR. See I. Mendoza and L. A. Bygrave, 'The Right Not to Be Subject to Automated Decisions Based on Profiling' in: T. Synodinou, P. Jougleux, C. Markou and T. Prastitou (eds.) *EU Internet Law: Regulation and Enforcement* (Dordrecht: Springer, 2017), pp. 77–98 at 90–91; Sartor and Lagioia (note 21), p. 60. The better view (at least *lex lata*) is that profiling is one of two alternative baseline criteria for the application of Article 22(1): see further L. A. Bygrave, 'Article 22: Automated Individual Decision-Making, Including Profiling' in: C. Kuner, L. A. Bygrave and C. Docksey (eds.) *The EU General Data Protection Regulation (GDPR): A Commentary* (Oxford: Oxford University Press, 2020), pp. 522–542, at 530. Article 4(4) GDPR defines 'profiling' as '*any form of automated processing* of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements' (emphasis added). In any case, it is likely that many if not most decisional systems caught by Article 22(1) will involve profiling as so defined.

[84] Bygrave (note 81), p. 249.

[85] See, especially, recital 71 GDPR.

involved, as well as the significance and the envisaged consequences of such processing for the data subject'.[86]

The provisions of Article 22 are intricate, complex and clumsily structured. They operate with multiple layers of qualifications, some of which involve cross-referencing beyond Article 22. One layer of qualifications inheres in Article 22(2), which provides that the right not to be subject to the decisions described in Article 22(1) does not exist in three alternative sets of circumstances: (a) the decision making is necessary for entering into or performing a contract with the data subject; (b) there is statutory authority for the decision making; or (c) the decision making is consented to by the data subject. Another layer inheres in Article 22(3), which basically states that, regardless of the exemptions of contract and consent in Article 22(2), the data subject must have 'at least' three further rights – the right to obtain human intervention, the right to express their point of view and the right to contest the decision – but only when the decision making is pursuant to contract or consent. On top of these layers comes that of Article 22(4), which places a qualified ban on automated decisions based on categories of especially sensitive personal data (as listed in Article 9 GDPR: e.g., data on health, philosophical beliefs or sexual orientation). Additionally, decisions based on 'systematic and extensive evaluation of personal aspects of natural persons' and which 'produce legal effects concerning the natural person or similarly significantly affect the natural person' are subject to *ex ante* 'data protection impact assessment' (Article 35(3)(a) GDPR). Such decisions need not be fully automated; they thus extend beyond those referred to in Article 22(1).[87]

In-depth analysis of the numerous facets of these provisions is well beyond the scope of this chapter;[88] the aim here is to highlight their basic thrust, along with principal points of controversy and uncertainty afflicting them. One point of controversy concerns the very need for Article 22. Some critics query its underlying premise that machine-based decisions are intrinsically problematic, and argue that this fails to do justice to the potential benefits of such decisions.[89] My own view is that the net cast by Article 22 is sufficiently refined as to capture only those decisions requiring special scrutiny – that is, decisions that have legal or similarly significant effects on data subjects. Thus, it is not targeting machine-based decisions per se.

More salient points of controversy and uncertainty concern the nature of the right in Article 22(1) and whether or not data subjects are to be provided with a right of *ex post* explanation of automated decisions affecting them (in addition to the three rights listed in Article 22(3)). Regarding the nature of the right in Article 22(1), there is widespread perception that it is

[86] The term 'data subject' denotes the natural/physical person to whom the data relate: see Article 4(1) GDPR. In the context of Article 22, the data subject can also be regarded as the decision subject (although the GDPR eschews the latter terminology).

[87] See too Article 29 Working Party, Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679 (WP 251rev.01, as last revised and adopted on 6 February 2018), p. 29 ('Article 35(3)(a) will apply in the case of decision making including profiling with legal or similarly significant effects that is *not* wholly automated, as well as solely automated decision making defined in Article 22(1)').

[88] For relatively comprehensive analyses, see, e.g., Mendoza and Bygrave (note 83); M. Brkan 'Do Algorithms Rule the World? Algorithmic Decision Making and Data Protection in the Framework of the GDPR and Beyond' (2019) 27, no. 2 *International Journal of Law and Information Technology* 91–121. https://doi.org/10.1093/ijlit/eay017; Bygrave (note 83).

[89] See, e.g., D. Kamarinou, C. Millard and J. Singh, 'Machine learning with personal data'. (2016) Queen Mary School of Law Legal Studies Research Paper No. 247/2016. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2865811, p. 22 ('machines may soon be able to overcome certain key limitations of human decision makers and provide us with decisions that are demonstrably fair. Indeed, it may already in some contexts make sense to replace the current model, whereby individuals can appeal to a human against a machine decision, with the reverse model whereby individuals would have a right to appeal to a machine against a decision made by a human').

essentially a prohibition masquerading as a right – in other words, that Article 22(1), despite its formulation, is really laying down a qualified ban on fully automated decisional processes, independently of data subjects' objections.[90] This view has influential supporters, in particular European Data Protection Authorities (DPAs),[91] and it clearly gives Article 22 greater traction over the deployment of automated decision making. However, the better view (at least *lex lata*) is that Article 22(1) provides a right to be exercised at the discretion of data subjects. This is consistent not only with the actual wording of Article 22(1) but also the GDPR's legislative history, and makes sense in light of other provisions of the GDPR.[92]

With respect to the existence of a right to *ex post* explanation, it has been argued that the GDPR does not operate with such a right, at least in a legally enforceable form.[93] The better view (both *lex lata* and *lex ferenda*) is that the right inheres in multiple parts of the Regulation. These are primarily the previously cited provisions of Articles 13(2)(f), 14(2)(g) and 15(1)(h), along with the right to contest a decision (Article 22(3)) – which is really a right of review and thus must involve the supply of reasons (at least if it is to be fair) – and the overarching requirement that personal data be processed 'fairly and in a transparent manner' (Article 5(1)(a)).[94] The right also arguably follows from the general accountability requirements to which data controllers[95] are subject pursuant to Articles 5(2) and 24. However, no such right exists under Article 11 LED with respect to the criminal justice sector.

Viewed as a whole from a fundamental rights perspective, the most valuable achievement of Article 22 is that it gives a data subject an unqualified right to demand proper human review of a fully automated decision involving processing of personal data, except where the decision does not have the requisite effects as specified in Article 22(1) or where the decision is pursuant to statute. Even if statutory authority kicks in, this might well also allow for a right of review as lawmakers must always provide for 'suitable measures to safeguard the data subject's rights and freedoms and legitimate interests' (Article 22(2)(b)). In many contexts – particularly those involving exercise of state authority over citizens – such measures necessitate a right of review, not just under data protection law but under laws dealing with human rights and administrative decision making more generally.

The GDPR also goes a long way in promoting cognitive sovereignty in relation to the decisional contexts it covers. In particular, the requirement that 'meaningful information' be provided about the 'logic' of fully automated decision making along with its 'significance and envisaged consequences … for the data subject' is far more exacting than a requirement of decisional transparency. Whereas the latter requirement on its own does not necessarily generate greater comprehensibility (and can be used as an instrument of obfuscation, particularly in

[90] As is the case under Article 11(1) LED, which expressly prohibits (subject to exceptions) such decision making in respect of the criminal justice sector.

[91] Article 29 Working Party (note 87), pp. 19–20.

[92] L. Tosoni, 'Right to Object to Automated Individual Decisions: Resolving the Ambiguity of Article 22(1) of the General Data Protection Regulation' (2021) 11, no. 2 *International Data Privacy Law*, 145–162. https://doi.org/10.1093/idpl/ipaa024.

[93] S. Wachter, B. Mittelstadt and L. Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7, no. 2 *International Data Privacy Law* 76–99. https://doi.org/10.1093/idpl/ipx005.

[94] Bygrave (note 81).

[95] The term 'controller' denotes the entity that determines – or co-determines – the means and purposes of data processing: see further Article 4(7) GDPR. This entity is also the decision maker for the purposes of Article 22.

relation to complex processes),[96] the former requirement signals a functional concern for explicability tailored to the particular cognitive needs of the data subject.[97] This is backed up by the stipulation that information provided to the data subject be 'concise, transparent, intelligible and easily accessible . . ., using clear and plain language' (Article 12(1) GDPR). In setting these exacting standards, the GDPR affords persons with relatively extensive capability to understand the functionality, impact, consequences and rationales of automated decision making. Indeed, Malgieri and Comandé convincingly argue that the GDPR mandates 'legibility-by-design': that is, a requirement to incorporate legibility (the term they use to denote comprehensibility) in the architecture of decisional processes at the outset of their design.[98] This requirement can be seen, in turn, as part and parcel of the more general 'data protection by design and by default' requirements in Article 25 GDPR.[99] It bolsters XAI efforts[100] and, if properly acted on, ought to ameliorate the potential conundrum facing decision makers who are obliged to explain the logic of an ML-enhanced decisional process but who struggle themselves to understand that logic.

As for the tension between the interest in cognitive sovereignty and the interest in protecting the commercially sacrosanct, the GDPR acknowledges that the right of data subjects to request information about the logic of automated decisional processes 'should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software' (recital 63; see too Article 23(1)(i)), but it also stipulates that these considerations 'should not' result in 'a refusal to provide all information to the data subject' (recital 63). Moreover, recitals 34 and 35 of the EU Directive on Trade Secrets[101] essentially state that the trade secrets protections laid down therein must not ride roughshod over the data protection rights of persons whose personal data are processed by a trade secret holder, in particular their information access rights. In light of these provisions, Malgieri and Comandé argue, in effect, that EU law ends up providing greater priority to the interest in cognitive sovereignty than the interest in safeguarding the commercially sacrosanct.[102] It is perhaps more

[96] As Pasquale (note 48), p. 8, remarks, 'transparency may simply provoke complexity that is as effective at defeating understanding as real or legal secrecy. [. . .] Transparency is not just an end in itself, but an interim step on the road to intelligibility'. Further on the limits of transparency as 'silver bullet', see, e.g., M. Ananny and K. Crawford, 'Seeing without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability' (2016) *New Media & Society*, December, 1–17. https://doi.org/10.1177/1461444816676645; L. Edwards and M. Veale, 'Slave to the Algorithm: Why a "Right to an Explanation" Is Probably Not the Remedy You Are Looking for' (2017) 16, no. 1 *Duke Law & Technology Review* 18–84.

[97] See e.g. A. D. Selbst and J. Powles, 'Meaningful Information and the Right to Explanation' (2017) 7, no. 4 *International Data Privacy Law* 233–242. https://doi.org/10.1093/idpl/ipx022; Article 29 Working Party (note 87), pp. 19–20.

[98] G. Malgieri and G. Comandé, 'Why a Right to Legibility of Automated Decision-Making Exists in the General Data Protection Regulation' (2017) 7, no. 4 *International Data Privacy Law* 243–265. https://doi.org/10.1093/idpl/ipx019.

[99] Further on the latter requirements, see, e.g., European Data Protection Board, Guidelines 4/2019 on Article 25: Data Protection by Design and by Default (version 2.0; 20 October 2020).

[100] See too M. Brkan and G. Bonnet (2020). 'Legal and Technical Feasibility of the GDPR's Quest for Explanation of Algorithmic Decisions: of Black Boxes, White Boxes and Fata Morganas' 11, no. 1 *European Journal of Risk Regulation* 18–50. https://doi.org/10.1017/err.2020.10, p. 19 ('The GDPR is . . . becoming increasingly important . . . for XAI researchers and algorithm developers, since the introduction of the legal requirement for understanding the logic and hence explanation of algorithmic decisions entails also the requirement to guarantee the practical feasibility of such explanations from a computer science perspective').

[101] Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure [2016] O.J. L 157/1.

[102] Malgieri and Comandé (note 98), pp. 263–264. See also G. Malgieri, *'Trade Secrets v Personal Data*: A Possible Solution for Balancing Rights' (2016) 6, no. 1 *International Data Privacy Law* 102–116, at 104–105. https://doi.org/10.1093/idpl/ipv030.

accurate to state that EU law does not permit the one interest to cancel out the other, and that compromises must be struck such that cognitive sovereignty is promoted to the greatest extent possible without unduly undermining protection for trade secrets or IPR.

In the latter regard, a variety of methods exist to facilitate some degree of comprehension of automated decisional logic without revealing the complete source-code of algorithms or otherwise unduly compromising their protection as trade secrets or IPR.[103] However, as Brkan and Bonnet demonstrate, these methods tend to fall short in providing an optimal granularity of explanation, at least for non-experts.[104] Particular types of decisional systems are especially challenging in this respect, with neural networks and 'multi-agent' systems as principal examples in point.[105]

These shortcomings are somewhat offset by a range of computational tools to promote algorithmic accountability independently of mechanisms directly geared to promoting cognitive sovereignty.[106] Additionally, we must be careful not to view cognitive sovereignty as always being the chief goal for data subjects who feel unjustly treated by algorithmic regulation or other decisional processes to which they are subject. Edwards and Veale observe that, in respect of many 'algorithmic "war stories" that strike a public nerve … what the data subject wants is not an explanation – but rather for the disclosure, decision or action simply not to have occurred'.[107] Explanations do not of themselves prevent unfairness, particularly when deeply ingrained structural imbalances of power exist between data subjects and decision makers. Thus, data subjects do not necessarily seek explanations as their primary remedy for unfair or unjust decisions; their primary remedy could well be utilising the rights afforded by Article 22. These rights go well beyond a concern for cognitive sovereignty.

The remedial power of Article 22, though, has weaknesses. Its application is neither clear nor straightforward owing to difficulties in working out precisely when a 'decision' has been made and what sort of effects the decision has on data subjects. Such difficulties are particularly acute in ML contexts.[108] The requirement that a decision be based 'solely' on an automated process is another stumbling block for application of Article 22; when the automated elements function as decisional *support* – that is, when there is a 'human in the loop' – Article 22 is irrelevant. The few court cases involving possible application of Article 22 or its antecedents have tended to result in findings that the decisional system at issue was not fully automated.[109] Nonetheless, it bears emphasis that the 'human in the loop' needs to take an active role in the decisional process; the person cannot slavishly comply with the recommendation of the algorithm(s) but must actively

[103] See, e.g., Brkan and Bonnet (note 100), 40–41; N. Diakopoulos, 'Accountability in Algorithmic Decision Making' (2016) 59, no. 2 *Communications of the ACM* 56–62, at pp. 57ff. https://doi.org/10.1145/2844110; S. Wachter, B. Mittelstadt and C. Russell, 'Counterfactual Explanations without Opening the Black-Box: Automated Decisions and the GDPR' (2018) 31, no. 2 *Harvard Journal of Law and Technology* 841–887. Available at https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.

[104] Brkan and Bonnet (note 100), pp. 37–38.

[105] A multi-agent system is composed of multiple decisional units, such as a network of connected, automated vehicles: Brkan and Bonnet (note 100), p. 38.

[106] J. A. Kroll et al., 'Accountable Algorithms' (2017) 165 *University of Pennsylvania Law Review* 633–705.

[107] Edwards and Veale (note 96), p. 42.

[108] Ibid., pp. 46–48. See also R. Binns and M. Veale, 'Is That Your Final Decision? Multi-stage Profiling, Selective Effects, and Article 22 of the GDPR' (2021) ipab020. https://doi.org/10.1093/idpl/ipab020.

[109] See judgment of 28 January 2014 by the German Federal Court of Justice in the so-called SCHUFA case concerning credit scoring (VI ZR 156/13); judgment of 11 February 2020 by the District Court of The Hague in the 'E-screener' case concerning a digital questionnaire used to evaluate the psychological conditions of gun license applicants (ECLI:NL:RBDHA:2020:1013); judgment of 18 December 2020 by the Austrian Federal Administrative Court concerning the Austrian Labour Market Service's use of the 'Arbeitsmarktchancen Assistenz-System' ('AMAS') for assessing the job prospects of persons seeking employment (ECLI:AT:BVWG:2020:W256.2235360.1.00).

consider its merits before reaching a decision.[110] This is especially important in light of empirical evidence of 'automation bias'.[111] Some scholars intimate, in effect, that such evidence means that many AI-supported decisions will indeed fall within the scope of Article 22(1).[112] In any case, a decisional process that is not fully automated will be caught by other provisions of the GDPR, as long as the process utilises personal data. Other regulatory codes may apply too, as elaborated further in this section.

The weaknesses of Article 22 as a tool for 'taming the machine' are augmented by limitations inherent in the scope and character of the GDPR itself. With respect to scope, the GDPR applies only to the processing of *personal* data; hence, decisional systems not involving the processing of such data fall outside its ambit. This limitation applies for data that relate to collective entities but cannot otherwise be readily linked to a particular identifiable, individual natural/physical person.[113] It reflects a general exclusion of 'aggregate' or 'group' data from the ambit of data protection law.[114] This undercuts the ability of such law to promote the cognitive sovereignty and related data protection interests of collective entities: a difficulty that has become increasingly problematic in an era where myriad group profiles are created and deployed, in part through the application of BDA.[115] However, it needs to be borne in mind that the definition of 'personal data' in the GDPR is expansive and that the use of BDA and related technologies engenders a situation where large swathes of 'aggregate' data that appear to be anonymous are able to be linked to specific individuals,[116] thus bringing the data within the GDPR's ambit.

As for the GDPR's character, there are several interlinked problems, particularly with respect to ML. First, the regulatory thrust of the GDPR – like other data protection laws – is directed less towards how information and thereby knowledge are generated than to how information is utilised, disseminated and stored (as units of personal data). The rules of data protection law do not engage directly with the processes involved in creating models, algorithms and other elements of inferential architecture that are critical to ML as an epistemic enterprise.[117] This is not to say that data protection law has no bearing on such processes. Its requirements that

[110] Bygrave (note 83), p. 21.

[111] See, e.g., L. J. Skitka, K. L. Mosier and M. Burdik, 'Does Automation Bias Decision-Making?' (1999) 51, no. 5 *International Journal of Human-Computer Studies* 991–1006. https://doi.org/10.1006/ijhc.1999.0252 (documenting an observed tendency of humans to commit 'errors of omission' (i.e., missing events when the computer fails to provide notification of them) and 'errors of commission' (i.e., following the computer's recommendation even when this contradicts humans' training and other reliable indicators).

[112] Sartor and Lagioia (note 21).

[113] Bygrave and Tosoni (note 76).

[114] However, where non-personal data are 'inextricably linked' with personal data (meaning that they cannot be processed separately of each other), the GDPR applies to both types of data: see Article 2(2) of Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union [2018] O.J. L 303/59.

[115] See further, e.g., A. Mantelero, 'Personal Data for Decisional Purposes in the Age of Analytics: From an Individual to a Collective Dimension of Data Protection' (2016) 32, no. 2 *Computer Law & Security Review* 238–255. https://doi.org/10.1016/j.clsr.2016.01.014; B. Mittelstadt, 'From Individual to Group Privacy in Big Data Analytics' (2017) 30, no. 4 *Philosophy and Technology* 475–494. https://doi.org/10.1007/s13347–017-0253-7.

[116] N. Purtova, 'The Law of Everything: Broad Concept of Personal Data and Future of EU Data Protection Law' (2018) 10, no. 1 *Law, Innovation and Technology* 40–81; M. Finck and F. Pallas, 'They Who Must Not Be Identified: Distinguishing Personal from Non-personal Data under the GDPR' (2020) 10, no. 1 *International Data Privacy Law* 11–36. https://doi.org/10.1093/idpl/ipz026.

[117] S. Wachter and B. Mittelstad, 'A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI' (2019) 2019, no. 2 *Columbia Business Law Review* 494–620; R. Gellert, 'Comparing Definitions of Data and Information in Data Protection Law and Machine Learning: A Useful Way Forward to Meaningfully Regulate Algorithms?' (2022) 16, no. 1 *Regulation & Governance* 156–176. https://doi.org/10.1111/rego.12349.

personal data be relevant and compatible in relation to the purposes for which they are used[118] signal a concern to ensure that data controllers duly reflect over the nature of the problems/tasks for which they process data, and over the quality (relevance, validity etc.) of the data they process to address those problems/tasks.[119] But the signal is far from strong, and, again, it does not apply to all data, only to personal data.

This brings us to a second problem with the GDPR's general character which is that its requirements – and those of data protection laws in general – are directed primarily at data *controllers* (i.e., entities that determine or co-determine the purposes and means of processing personal data) and *processors* (i.e., entities that process personal data on behalf of controllers).[120] These actors will not necessarily be engaged in the basic design and development of information systems, including the creation of models, algorithms etc. integral to ML-based decisional processes. This not only undermines the GDPR's aim that its norms become 'hard-wired', as it were, into information systems architecture[121] but also points to a general weakness in the ability of data protection law to engage with the engineers and computer scientists who are directly involved in designing and constructing that architecture. This weakness is exacerbated by a third 'character' problem, which is that the GDPR is formulated in a dense, often vague legalese that communicates poorly with the engineering community and with others lacking bespoke legal expertise.

Some critics claim that there is another major 'character' flaw, which is that the 'fair information practice principles' (FIPPs) at the core of the GDPR (and other data protection laws) are fundamentally out of tune with the realities of current data-processing practices. Zarsky, for instance, states bluntly that '[t]he GDPR's provisions are [...] *incompatible* with the data environment that the availability of Big Data generates'.[122] Others argue that data protection law in the face of BD is a 'largely useless Maginot Line',[123] while yet others have sounded its death knell.[124] My own view is more sanguine. Certainly some of the fundamental assumptions of the FIPPs are in tension with the logic of BDA. For example, the requirement that personal data be processed for predefined specific purposes[125] is challenging to reconcile with the fact that BDA is often initiated without very specific objectives, and the requirement that the processing of personal data be transparent and explainable is challenging to implement when many ML-based processing operations are inherently opaque. Yet, these tensions do not amount to a frontal collision that stops BDA in its tracks or prevents innovation more generally.[126]

At the same time, the FIPPs are sufficiently flexible to avoid being rapidly marginalised by BDA or other technological developments. The principle that personal data shall be processed

[118] See especially Articles 5(1)(b) and (c) GDPR.
[119] Bygrave (note 58), pp. 148–150, 337.
[120] See Articles 4(7) and 4(8) GDPR.
[121] See Article 25 combined with Articles 28(1) and 24. See further L. A. Bygrave, 'Data Protection by Design and by Default: Deciphering the EU's Legislative Requirements' (2017) 4, no. 2 *Oslo Law Review* 105–120. https://doi.org/10.18261/issn.2387-3299-2017-02-03, p. 118.
[122] T. Zarsky, 'Incompatible: The GDPR in the Age of Big Data' (2017) 47, no. 4 *Seton Hall Law Review* 995–1020, p. 996.
[123] Mayer-Schönberger and Cukier (note 19), pp. 15–16.
[124] B. J. Koops, 'The Trouble with Europên Data Protection Law' (2014) 4, no. 4 *International Data Privacy Law* 250–261. https://doi.org/10.1093/idpl/ipu023.
[125] See, e.g., Article 5(1)(b) GDPR.
[126] See too, e.g., N. Forgó, S. Hänold and B. Schütze, 'The Principle of Purpose Limitation and Big Data' in: M. Corrales, M. Fenwick and N. Forgó (eds.) *New Technology, Big Data and the Law* (Dordrecht: Springer; 2017), pp. 17–42; Sartor and Lagioia (note 21), pp. 4ff.

'fairly'[127] is especially important in this regard. Some scholars query the utility of the fairness criterion, suggesting, in effect, that it is not much more than a proxy for transparency.[128] In my view, fairness has independent work to do and potentially brings with it a host of normative parameters over and above transparency, such as protection from unwarranted discrimination, power imbalance and risk or vulnerability. This is also the view of DPAs,[129] some of which have begun to apply the fairness criterion in concrete disputes. An instructive example is the recent draft decision of Norway's DPA to disallow use by the International Baccalaureate Organization (IBO) of a covert profiling algorithm to determine the final grades of students who were unable to sit exams in 2020 owing to the COVID-19 pandemic: in the DPA's opinion, the IBO's grading system and failure to disclose fully the algorithmic logic involved was in breach of, inter alia, the fairness criterion in Article 5(1)(a) GDPR.[130]

Finally, it is important to note that the GDPR is but one element of a multifaceted data protection regime. The 'weapons arsenal' the regime can bring to bear on BDA and other data-processing practices extends to provisions in basic human rights instruments, such as Articles 7 and 8 of the EU Charter of Fundamental Rights and Article 8 of the European Convention on Human Rights and Fundamental Freedoms (ECHR). The breadth and flexibility of this arsenal are neatly illustrated by the recent judgment of the District Court of The Hague in the 'SyRI' case.[131] This dealt with the legality of an automated risk indication system ('Systeem Risico Indicatie' ('SyRI')) established by the Dutch government to combat fraud in the areas of tax and social security. The system allowed for the combination of structured datasets kept by government agencies to produce risk reports indicating that a person is worthy of investigation for possible fraud or other unlawful behaviour. The system, including the algorithms applied, was far from transparent. The court struck down the legislation governing its use for violating Article 8(2) ECHR. This was because of the state's failure to strike a 'fair balance' between the legislation's objectives and the interference it caused with the right to respect for private life.[132] Drawing on case law of the European Court of Human Rights (ECtHR), the court held that a state which is at the forefront of employing new technologies that can interfere extensively with the private lives of those to whom the technologies are applied bears a 'special responsibility' to strike such a balance.[133] In this case, the lack of publicly available information about key aspects of the system's logic and mechanics, combined with the large amounts and numerous categories of personal data processed by the system, along with deficient review mechanisms, meant that the requisite balance was not struck. In particular, the court highlighted that the system's opacity was not just a problem for the data subjects but also for the system's own ability to demonstrate its proportionality.[134] The court declined to reach a definitive view as to whether or not SyRI met

[127] See Article 5(1)(a) GDPR.

[128] S. Eskens, *Profiling the European Citizen in the Internet of Things: How Will the General Data Protection Regulation Apply to this Form of Personal Data Processing, and How Should It?* (29 February 2016; Master's thesis, University of Amsterdam). Available at http://dx.doi.org/10.2139/ssrn.2752010, p. 27 (fn. 125); Wachter and Mittelstad (note 117).

[129] European Data Protection Board (note 99), para 70.

[130] Draft decision of 7 August 2020 in Case 20/03087.

[131] Rechtbank Den Haag, 5 February 2020 (ECLI:NL:RBDHA:2020:1878).

[132] Ibid., paras. 6.80 and 6.83.

[133] Ibid., para. 6.84ff. See *S and Marper* v. *United Kingdom*, ECtHR 4 December 2008, application nos. 30562/04 and 30566/04, para. 112 ('The Court considers that any State claiming a pioneer role in the development of new technologies bears special responsibility for striking the right balance in this regard').

[134] Ibid., para. 6.95.

the conditions for applying Article 22 GDPR, noting that the issue was irrelevant to the review of the system's legality pursuant to Article 8 ECHR.[135]

13.5 CONCLUSION: A TURN TO 'FIRST PRINCIPLES'?

The GDPR goes a long way in ensuring, on paper at least, that ML-enhanced and other automated decisional systems respect our cognitive sovereignty and related data protection interests, and that such systems must potentially involve an informed 'human in the loop'. At first glance, this capability seems primarily to flow from those provisions of the GDPR that deal directly with automated decision making: that is, Article 22, along with Articles 13(2)(f), 14(2)(g) and 15(1)(h). Their antecedents in the 1995 Data Protection Directive were predominantly left slumbering.[136] This is a fate they are unlikely to share. It remains to be seen, though, whether the rights they provide will be actively pursued by large numbers of us in our capacity as data/decision subjects. Much will depend on the extent to which DPAs and, secondarily, courts vigorously promote them. Moreover, the traction of the ideals these rights seek to safeguard will depend on the degree to which regulators provide clear, timely guidance on what the provisions require of those responsible for constructing or deploying automated decisional systems. The message flowing from the GDPR itself is in many respects cryptic, especially for the engineering community. Hence, even though the message aligns with the thrust of XAI endeavours, one cannot simply assume that it will play a large practical role in animating them.

As shown in the previous section, the provisions are part of a much larger regulatory 'arsenal' comprising more generally worded norms that can do most, if not all, of the work that the former are intended to do. Reliance by DPAs and courts on the latter can avoid having to address many of the tricky 'boundary' issues afflicting application of Article 22 (such as whether a 'decision' has been made and whether the decision is based 'solely' on automated processing). This creates doubts over the long-term utility of Article 22 in regulating ML-enhanced or other automated decision making. In other words, we may duly ask whether its practical relevance will be undermined by a turn to 'first principles' utilising criteria based on notions of proportionality, balance and fairness. At the same time, another troubling question raises its head: is there not a somewhat paradoxical risk that reliance on the latter criteria produces what could be termed 'grey-box' decision making: that is, somewhat cryptic decisions grounded in woolly interest-balancing processes informed by relatively subjective notions of propriety?

In my view, the provisions of Article 22 will remain useful for helping to flesh out the substance of these otherwise relatively diffuse criteria, even if the former do not specify the outer limits of the latter. Moreover, the latter do not present unchartered territory: a rich long line of jurisprudence applies and elucidates their key characteristics.[137] Nonetheless, the

[135] Ibid., para. 6.60. The Dutch government decided not to appeal the decision but to scrap SyRI and build a new control system with improved transparency and review mechanisms: see T. van Ark, Letter of 23 April 2020 to the President of the House of Representatives by the State Secretary for Social Affairs and Employment, Tamara van Ark, on a court judgment regarding SyRI. Available at: www.rijksoverheid.nl/ministeries/ministerie-van-sociale-zaken-en-werkgelegenheid/documenten/publicaties/2020/04/23/vertaling-kamerbrief-naar-aanleiding-van-vonnis-rechter-inzake-syri.

[136] Bygrave (note 81), p. 250.

[137] For relatively comprehensive presentations, see e.g. A. Barak, *Proportionality: Constitutional Rights and Their Limitations* (Cambridge: Cambridge University Press, 2012), Part II; A. Stone Sweet and J. Mathews, 'Proportionality Balancing and Global Constitutionalism' (2008) 47 *Columbia Journal of Transnational Law* 68–149.

mechanics of proportionality assessment are multiplex, complicated and frequently disputed; its outcomes are accordingly often challenging to predict or comprehend.[138] Thus, to mitigate the 'grey-box' risk, it is incumbent on courts and DPAs to explain clearly how they utilise 'first principles', also in assessments of opaque automated decisional systems. Not only does our sense of justice demand this, but our interest in cognitive sovereignty demands it too.

[138] The case, for instance, with the 'margin of appreciation' doctrine in ECtHR jurisprudence (see further L. Wildhaber, A. Hjartarson and S. Donnelly, 'No Consensus on Consensus? The Practice of the European Court of Human Rights' (2013) 33, no. 7 *Human Rights Law Journal* 248–263) and with the role of the doctrine on respecting the 'essence' of a fundamental right in CJEU jurisprudence (see further M. Brkan, 'The Essence of the Fundamental Rights to Privacy and Data Protection: Finding the Way through the Maze of the CJEU's Constitutional Reasoning' (2019) 20 *German Law Journal* 864–883).