

## 2

# *The Rule of Algorithm and the Rule of Law*

JOHN TASIOULAS\*

ATHENIAN – And you realise, don't you, that the people who fall sick in our cities may be slaves or free-born? And that it is the slave-doctors who for the most part treat the slaves, either dashing round the city or sitting in their surgeries? None of these doctors gives any explanation of the particular disease of any particular slave – or listens to one; all they do is prescribe the treatment as they see fit, on the basis of trial and error – but with all the arrogance of a tyrant, as if they had exact knowledge. Then they're up and off again, to the next suffering slave, and in this way they give their masters a breathing-space in their caring for the sick.

The free-born doctor spends most of his time treating and keeping an eye on the diseases of the free-born. He investigates the origin of the disease, in the light of his study of the natural order, taking the patient himself and his friends into partnership. This allows him both to learn from those who are sick, and at the same time to teach the invalid himself, to the best of his ability; and he prescribes no treatment without first getting the patient's consent. Only then, and all the time using his powers of persuasion to keep the patient cooperative, does he attempt to complete the task of bringing him back to health. Is a doctor who heals in this way a better doctor? Or the other way? Likewise a trainer who trains in this way? He has one single ability. Should he get it to complete its exercise by this dual method, or in the simple way – the less good of the two, and the one which makes the patient more hostile?

CLEINIAS – The dual approach, my friend, is by far the better.<sup>1</sup>

\* This chapter has its origins in a lecture I delivered at the University of Vienna on 15 October 2021. I am grateful to Professor Alexander Somek for the kind invitation to deliver his lecture and for his gracious hospitality. For helpful comments on a subsequent draft I am grateful to Jeremias Adams-Prassl, Roger Brownsword, Grant Lamond, Hélène Landemore, Rebecca Lowe, Aislinn Kelly-Lyth, Onora O'Neill, Elizabeth Renieris, Divya Siddarth, Richard Susskind, Adrian Vermeule, Mike Woolridge and John Zerilli.

<sup>1</sup>Plato, *The Laws*, M Schofield (ed) (Cambridge, Cambridge University Press, 2016) 163–64 (720b-e).

## I. ARTIFICIAL INTELLIGENCE (AI) AND LEGAL ADJUDICATION

WE ARE LIVING in the midst of a heady AI spring. It is marked by a profusion of hopes, dreams, and fantasies about the potential of AI-based technologies to make unprecedented advances in furthering our personal and collective goods. The material underpinnings of this optimism are three familiar, large-scale developments that have unfolded in recent decades: an exponential growth in computing power; the availability of huge amounts of digital data; and the emergence of increasingly powerful algorithms that, by means of the science of machine learning, draw on the first two developments to yield remarkable improvements in the ability of automated systems to perform tasks that ordinarily require intelligence when done by humans. These tasks range from cancer diagnosis and facial recognition to the ranking of job applicants' resumes and assessments of the risk that a criminally accused would commit an offence were they to be released on bail. Sometimes AI applications seemingly outperform humans, being able to discern complex patterns in the data that no human can feasibly detect, and to use those patterns as a basis for classifications, predictions or decisions in new cases.

The law is an important focus for many of these AI-driven aspirations. The current global situation regarding access to justice through law is a dire one. The Organisation for Economic Co-operation and Development's (OECD) research has found that only 46 per cent of the world's population lives under the protection of the law, meanwhile the backlog of cases in court systems across the globe includes 30 million cases in India and 100 million cases in Brazil.<sup>2</sup> Legal justice is one of the most precious societal goods but, in democracies, it is rather more challenging to muster electoral support for increased public expenditure on it than on matters such as health and security. Meanwhile, in authoritarian societies there are strong vested interests militating against its fair distribution. Hence the political allure of the idea that AI, and digital technology more generally, might be able to deliver legal services of comparable, or even superior, quality to those provided by humans, but much more rapidly and at a significantly reduced cost, and perhaps without the support or even involvement of governments.

The various potential intersections of AI technology with law is a vast topic. To make it somewhat more tractable, I am going to focus quite narrowly on one aspect of a legal system, that of adjudication of the kind undertaken by appellate judges, which is essentially a matter of rendering a decision in accordance with the law with respect to a given set of facts. And I will concentrate on one key value such adjudication should exemplify, which is compliance with the rule of law. Can AI help us better achieve the rule of law ideal in legal adjudication? To bring out the values at stake more forcefully, I will focus on the proposal that AI adjudicative tools replace, rather than merely assist, human judges, that is,

<sup>2</sup>R Susskind, *Online Courts and the Future of Justice* (Oxford, Oxford University Press, 2019) ch 28.

that they render legally binding decisions in particular cases submitted to them. Notice that I am not broaching here the more radical possibility of AI tools displacing governance through law by means of what Roger Brownsword has called ‘technological management’. The latter technique operates not by subjecting people to rules – which presupposes a choice on their part regarding compliance with those rules – but by using technology to channel people’s behaviour in ways that effectively preclude certain choices, for example, by disabling someone’s car if they have missed a payment on it. By eliminating the possibility of human choice technological management poses special challenges to the value of human dignity that will not be discussed here.<sup>3</sup>

Of course, this optimism about the prospects for AI-driven legal adjudication is not universally shared. Lord Sales, a current judge of the UK Supreme Court, has spoken of the widespread perception of AI as ‘fixed and remorseless, an infernal machine’, one that attracts three kinds of fears when deployed in the context of legal adjudication. In its remote and dehumanised character, it eludes democratic accountability; in its inability to modify the rigid application of law to avoid injustice in particular cases it fails to exercise equity;<sup>4</sup> and in its insensitivity to compassionate grounds for leniency it is deaf to pleas for mercy.<sup>5</sup> These are familiar worries. But the bearing of democratic accountability, and the ameliorative values of equity and mercy, on legal adjudication are deeply contested matters. Some have even questioned the relevance of these values to legal adjudication: Kant, for example, described equity as ‘a mute divinity that must not be heard’, and deemed the extension of mercy to offenders as the highest injustice except for narrowly restricted cases of *lèse majesté*. His scepticism is broadly shared by philosophers as different as Bentham and Hegel.<sup>6</sup>

Additionally, insofar as democracy, equity, and mercy are properly conceived as distinct from the rule of law, I set them aside, while acknowledging that they figure in a comprehensive account of good legal adjudication.<sup>7</sup> I will focus instead on the case for AI-based legal adjudication in relation to what looks like

<sup>3</sup> See R Brownsword, ‘Technological Management and the Rule of Law’ (2016) 8 *Law, Innovation and Technology* 100.

<sup>4</sup> But on this point, see Susskind (n 2) 289–90, on how machine learning algorithms, which can improve their performance on the basis of experience, can ‘may well generate output that appears creative and fresh’.

<sup>5</sup> Lord Sales, ‘Algorithms, Artificial Intelligence and the Law’ (The Sir Henry Brooke Lecture, 12 November 2019).

<sup>6</sup> Kant, *The Metaphysics of Morals*, L Denis (ed) (Cambridge, Cambridge University Press, 2017) 30. For a historical discussion of scepticism about mercy in the Kantian and utilitarian traditions, see A Tuckness and JM Parrish, *The Decline of Mercy in Public Life* (Cambridge, Cambridge University Press, 2014) chs 7–8.

<sup>7</sup> I have elsewhere defended a strong role for the values of equity and mercy in legal adjudication in a series of articles, including J Tasioulas, ‘The Paradox of Equity’ (1996) 55 *Cambridge Law Journal* 456; J Tasioulas ‘Mercy’ (2003) 103 *Proceedings of the Aristotelian Society* 101; and J Tasioulas, ‘Where is the Love? The Topography of Mercy’, in R Cruft, MH Kramer and M Reiff (eds), *Crime, Punishment and Responsibility: The Jurisprudence of Antony Duff* (Oxford, Oxford University Press, 2011) 37–53.

perhaps the most promising adjudicative value it can help us realise, that of the rule of law itself. Interpreted in the ‘thin’ way that I go on to endorse, the rule of law highlights formal values such as consistency in following rules, and treating like cases alike, which are often poorly satisfied in a system subject to the vagaries of fallible human judgment. Accordingly, my focus in this chapter is in line with the observation made by Richard Re and Alicia Solow-Niederman that ‘[t]he main strengths of AI adjudication are two hallmarks of codified justice: efficiency (or elimination of waste) and uniformity (or elimination of bias and arbitrariness)’.<sup>8</sup> However, my aim will be to raise doubts about the extent to which AI-driven adjudication can help us to realise even its most promising candidate adjudicative value – the kind of consistency in adjudication enjoined by the rule of law.

## II. THE RULE OF LAW – A THIN AND PLURALISTIC VIEW

What, then, does the rule of law require? The proper characterisation of the rule of law’s content is a matter of controversy. The views in play range from so-called ‘thin’ conceptions which restrict its content to certain formal-procedural desiderata, to ‘thick’ conceptions which also embrace a variety of substantive demands, such as democracy or the protection of human rights. The apogee of the thick approach, or else its nadir, depending on your viewpoint, is represented by the work of another eminent UK judge, Lord Bingham, in his assertion that the rule of law is ‘the nearest we’re likely to come to a universal secular religion’, an interpretation that encompasses pretty much all the main qualities of a good legal system.<sup>9</sup> Other, more circumspect defenders of a thick conception, tie the idea of the rule of law to the idea of legitimate law (Deryck Beylveid and Roger Brownsword) or to law that upholds individual moral rights (Ronald Dworkin).<sup>10</sup>

By contrast, the thin conception, which I favour, has been defended, in various ways, by theorists such as F A Hayek, Lon Fuller, Joseph Raz, John Finnis, and Cass Sunstein.<sup>11</sup> It identifies a series of formal-procedural desiderata

<sup>8</sup> R Re and A Solow-Niederman, ‘Developing Artificially Intelligent Justice’ (2019) 22 *Stanford Technology Law Review* 240, 255.

<sup>9</sup> See T Bingham, *The Rule of Law* (The Royal Society for Arts, Manufactures and Commerce, 10 March 2010) [www.youtube.com/watch?time\\_continue=9&v=X1MCCGD2TeM](http://www.youtube.com/watch?time_continue=9&v=X1MCCGD2TeM), but more fully (and circumspectly) developed in T Bingham, *The Rule of Law* (London, Allen Lane, 2010).

<sup>10</sup> D Beylveid and R Brownsword, *Law as a Moral Judgment* (London, Sweet & Maxwell, 1986) chs 7–9; R M Dworkin, *A Matter of Principle* (Cambridge, Harvard University Press, 1985) 11–12.

<sup>11</sup> FA Hayek, *The Road to Serfdom* (first published 1944, Chicago, University of Chicago Press, 1972); LL Fuller, *The Morality of Law*, revised edn (New Haven, Yale University Press, 1965) ch 2; J Raz, ‘The Rule of Law and its Virtue’, in J Raz (ed), *The Authority of Law* (Oxford, Oxford University Press, 1979); JM Finnis, *Natural Law and Natural Rights*, 2nd edn (Oxford, Oxford University Press, 2011) 270–71; CR Sunstein, *Legal Reasoning and Political Conflict*, 2nd edn (Oxford, Oxford University Press, 2018) 119–122.

pertaining to the enactment, adjudication, and enforcement of law. For present purposes we can simply refer to Fuller's recitation of eight principles of legality: laws must be framed as *general* rules, they must be *promulgated* to their subjects in advance, they must be *prospective* rather than retroactive in effect, they must be *clear, non-contradictory* and only require what is *possible* of their subjects, they must be relatively *stable* over time rather than constantly in flux, and there must be *congruence* between official conduct and the law.<sup>12</sup>

For the purposes of this chapter, I proceed on the basis of the thin conception of the rule of law for two main reasons. First, in terms of the dialectical context of my argument, the thin conception is the one that offers AI-driven technology its best shot at helping us to deliver the rule of law. The more we front-load the rule of law with additional – especially substantive – ideals the higher the bar it sets, and the bleaker the prospects are of AI tools serving as effective means of realising the rule of law. The desiderata captured by the thin conception are important ones, even if they don't come near to exhausting all the values that make for good legal systems. In addition, most proponents of thicker conceptions of the rule of law will include, as part of their overall account of its demands, the formal-procedural requirements that the thin conception identifies as exhaustive of those demands.

But there is also a more general, and principled, case for the thin conception. I have elsewhere set out an argument to the effect that this conception enables us to meet a two-pronged methodological demand: on the one hand, the pluralistic demand of elaborating the rule of law as one legally-relevant value among others; on the other hand, the demand that the desiderata flowing from the rule of law exhibit sufficient coherence or unity to justify bringing them together under one heading.<sup>13</sup> My contention is that thin theories do better in complying with this dual demand: according to these theories, the rule of law is one legally relevant value among others, distinct from goodness, justice, rights, legitimacy, and democracy; yet its various desiderata are unified by their shared character as procedural-formal. Thick theories, by contrast, tend to fall at the pluralistic hurdle, identifying the rule of law with the law's possession of some other value, such as justice or legitimacy or goodness. Given that we already have the concepts of justice, legitimacy, and goodness on hand, why collapse the putatively distinct idea of the rule of law into them?

### III. AI AS A VEHICLE FOR (BETTER) REALISING THE RULE OF LAW

The idea that AI can play an important role in legal adjudication, by delivering legal justice more efficiently without unduly sacrificing its quality, has by now

<sup>12</sup>See Fuller (n 11) ch 2.

<sup>13</sup>See J Tasioulas, 'The Rule of Law', in J Tasioulas (ed), *The Cambridge Companion to the Philosophy of Law* (Cambridge, Cambridge University Press, 2020) 117.

become familiar. I will concentrate on the rule of law dimension of this idea, and in particular on Fuller's eighth desideratum, that of congruence between the law and official conduct. The need for congruence follows from the fact that the law will not be able properly to achieve its action-guiding role for its subjects if the legal standards announced to those subjects are not those that are followed by officials in reaching their decisions.<sup>14</sup> The idea is that AI-generated decisions will be congruent with the law to at least the same degree as human decisions, while also making significant efficiency gains. Such decisions will be the outputs of algorithms – mechanical procedures that can be executed in a finite number of steps without any need for judgment, discretion or creativity.<sup>15</sup> Indeed, the implication is that they may provide for *even more* congruence than human decisions, given that we may realistically hope to design AI systems that are not afflicted (to the same degree) with certain defects that often subvert human officials' conformity with the law's requirements. These defects include myriad cognitive biases (eg, biases related to availability, anchors, affect heuristics, optimism, presentism, etc), ignorance, corruption, cowardice, prejudice, tiredness, and so on.

Let me now give some examples of proponents of this kind of view. Eugene Volokh has argued in favour of AI tools replacing judges, in principle, provided they can pass a legal version of the Turing test, such that the judgments they generate are able to persuade a panel of expert lawyers of their correctness at least to the same degree as judgments written by human judges.<sup>16</sup> Volokh's approach is in keeping with his mantras, 'consider the output, not the method' and 'what matters is the result, not the process'.<sup>17</sup> Given that the test of output is centred on the expectations of the panel of legal experts, one important expectation such experts will have is congruence between judicial decisions and the applicable law. Now, Volokh acknowledges that the process by which an AI adjudicative tool works to satisfy this legal Turing test may be totally unlike that of a human judge. So, to take an AI tool operating on the basis of deep learning, this will involve multiple layers of connected nodes, an activation function that determines the output of each node, and complex matrices of weights linking the nodes that encode the decision-making of the system, with a learning algorithm that recalibrates the relations among the nodes over time. However,

<sup>14</sup> Fuller calls congruence 'the most complex of all the desiderata that make up the internal morality of law', and says that it 'may be destroyed or impaired in a great variety of ways: mistaken interpretation, inaccessibility of the law, lack of insight into what is required to maintain the integrity of a legal system, bribery, prejudice, indifference, stupidity, and the drive toward personal power', Fuller (n 11) 81.

<sup>15</sup> 'An algorithm is a finite procedure, written in a fixed symbolic vocabulary, governed by precise instructions, moving in discrete steps ... whose execution requires no insight, cleverness, intuition, intelligence or perspicuity, and that sooner or later comes to an end.' D Berlinski, *The Advent of the Algorithm: The 300 Year Journey from an Idea to the Computer* (New York, Harcourt, 2000) xviii.

<sup>16</sup> E Volokh, 'Chief Justice Robots' (2019) 68 *Duke Law Journal* 1, 135.

<sup>17</sup> *ibid* 189.

the nodes will not be interpretable as the kinds of considerations (eg, legal rules, principles) that a competent human legal reasoner standardly takes into account in reaching a legal judgment. All this means that it is highly questionable that such an automated system operates in a way that can be accurately described as ‘following legal rules’. In response, Volokh stresses that what matters is not whether the system *complies* with legal rules, in the sense of taking them into account and being guided by them in the process of reaching a decision, so long as the decisions it yields *conform* to those rules to a sufficient degree:

The question is not whether an *AI judge* actually follows rules at some deep level; the question is whether an *AI judge’s opinions* persuade observers who expect opinions to be consistent with the legal rules. Rule-following is as rule-following does.<sup>18</sup>

A similar view is advanced by Richard Susskind. He emphasises the general significance of ‘outcome-thinking’, which ‘urges us to focus not on *how* humans do what they do, but on the outputs and benefits they bring’:

In the context of AI, this inclines us to consider whether machines can deliver decisions at the standard of human judges or higher, not by replicating the way that judges think and reason but by using their own distinctive capabilities (brute processing power, vast amount of data, remarkable algorithms).<sup>19</sup>

To fixate on the radical differences in the *ways* in which AI adjudicative tools achieve these valuable outcomes as compared to humans is, for Susskind, to fall victim to the ‘AI fallacy’, which uncritically assumes that ‘the only way for machines to do the work of human beings is for them somehow to mimic or copy the way that humans go about their business’.<sup>20</sup>

Another prominent example is Cass Sunstein, who believes that AI tools can help us avoid not only bias, but also noise, both of which impair the congruence between official decisions and applicable rules. As Sunstein puts it: ‘Noise consists in unwanted variability in judgments; by contrast bias consists of any systematic error that inclines people’s judgments in a particular direction.’<sup>21</sup> Here, I will focus on noise. In their recent book, Sunstein and his co-authors Daniel Kahneman and Olivier Sibony distinguish three forms of noise: (a) intra-personally, there is ‘occasion’ noise, where the same judge is influenced by irrelevant considerations in particular cases, for example, defeat suffered by his or her football team at the weekend; (b) inter-personally, there is ‘level’ noise where judgments in the system as a whole exhibit unwanted variability, for example, because some judges are on average more severe than others; and (c) there is also ‘pattern’ noise where the variability across judges does not take a

<sup>18</sup> *ibid* 161. The use of the comply/conform terminology is mine.

<sup>19</sup> Susskind (n 2) 280.

<sup>20</sup> *ibid*.

<sup>21</sup> CR Sunstein, ‘Governing by Algorithm? No Noise and (Potentially) Less Bias’ (2022) 71 *Duke Law Journal* 1, 175 and 178.

general form, but where instead there are various patterns of variability among judges, for example, some judges are severe in sentencing racial minorities but lenient towards middle class offenders, and vice versa.<sup>22</sup>

Now, there are potentially two kinds of ‘noise’, or ‘unwanted variability’, that contravene the desideratum of congruence: (a) where some decisions are not in line with the applicable law, either in the decisions of a particular judge or in the decisions made by various judges; and (b) where the law is being applied according to its terms, but so as to yield different results for identically placed litigants, whether in the case of the same judge or across the decisions of different judges. The latter situation can arise when the law itself leaves open a range of possible outcomes, for example, a bounded range of sentences of differing severity. This can be owing to inherent vagueness or incommensurability in the legally applicable considerations. The first type of noise is clearly inconsistent with the rule of law, even if in some cases (eg, the exercise of equitable judgment or mercy) it may be all-things-considered justified. The second is not obviously incompatible with the rule of law. Rather, it might be interpreted as violating a norm of formal justice, ‘treat like cases alike, unlike cases differently’, within an overall framework of respecting the rule of law. It is a delicate question whether one should incorporate that desideratum of formal justice within the rule of law and, in particular, as part of the requirement of congruence. Intuitions differ here over whether it is a moral defect that one criminally accused, for example, receives a lighter sentence than another accused whose case is indistinguishable from the former’s in all relevant respects, if both sentences taken individually fall within the legally (and morally) permissible range of sentences. But, given the strong association of the rule of law both with curbing arbitrary power and with formal-procedural requirements, we can treat it as a rule of law consideration for the sake of this discussion.

Now, Sunstein’s point is that, whatever the prospects of curing algorithms of biases, algorithms are nonetheless inherently silent or non-noisy.<sup>23</sup> At the same time, as he admits, the elimination of noise does not guarantee correctness of decision. An algorithm may be noiselessly yet persistently wrong because it is completely biased. The idea, however, is that the elimination of noise always prevents a morally undesirable form of unequal treatment, and typically helps advance other values, such as correctness. An egregious example of ‘noise’ offered by Sunstein and his co-authors pertains to the US asylum system. A study revealed that applications randomly allocated to different judges

<sup>22</sup>D Kahneman, O Sibony and CR Sunstein, *Noise: A Flaw in Human Judgment* (London, William Collins, 2021) 365–67.

<sup>23</sup>Sunstein (n 21) 185. But for scepticism about whether there can be true algorithms in the domain of action, providing fully determinate instructions as to their own enactment, see OS O’Neill, *From Principles to Practice: Normativity and Judgement in Ethics and Politics* (Cambridge, Cambridge University Press, 2018) 168–69. I bracket this important worry for the purposes of this discussion.

yielded admission rates of five per cent in the case of one judge and, at the other extreme, 88 per cent in the case of another judge.<sup>24</sup> By contrast, Sunstein writes:

If an applicant seeks asylum, the algorithm will offer the same answer whether it is Monday or Wednesday or January or June. Someone whose asylum application follows five successful applications will not be treated differently from someone whose application follows five unsuccessful applications. There is no occasion noise because the occasion cannot, and does not, matter. And because the level is the same across applications, there is no level noise. For the same reason algorithms cannot, and will not, display pattern noise. An algorithm with identical source code will not produce a different result in identical cases.<sup>25</sup>

Examples such as that of America's 'refugee roulette' give weight to the assertion by Sunstein and his co-authors that noise often involves great unfairness – an excruciating sense that decisions severely affecting people's life prospects exhibit unacceptable variability. And while Sunstein and his co-authors refrain from going so far as to propose the addition of a new right to the Universal Declaration of Human Rights – a novel form of the 'right to silence', as it were – they nonetheless insist that 'in some cases, noise can be counted as a rights violation, and in general, legal systems all over the world should be making much greater efforts to control noise'.<sup>26</sup>

#### IV. OUTPUT-BASED RESERVATIONS

One response to the arguments surveyed above is to contest them on their own 'output-based' terms. To begin with, it is scarcely to be taken for granted that AI tools will in fact be as reliable as human beings in identifying relevant legal standards, interpreting these standards correctly in the context of a given matrix of facts, and applying that interpretation to yield a legally sound decision in any given case. AI tools are subject to defects in functionality, such as manifesting bias owing to the unrepresentative data on which their algorithms are trained, or lacking the ability to extrapolate existing rules to new and unanticipated factual situations.<sup>27</sup> Indeed, no AI adjudicative tool in existence has ever passed anything like Volokh's judicial analogue of the Turing Test. As the authors of an important recent study have pointed out, in the absence of strong empirical confirmation of the capacities of AI adjudicative tools – such as that which would be furnished by a randomised controlled trial comparing human

<sup>24</sup> Kahneman, Sibony and Sunstein (n 22) 5–6.

<sup>25</sup> Sunstein (n 21) 185.

<sup>26</sup> Kahneman, Sibony and Sunstein (n 22) 359–60.

<sup>27</sup> For trenchant criticism of the use of AI tools in the legal system, with a special focus on their presupposed 'utilitarian predisposition' which is at odds with individual rights, see KB Forrest, *When Machines Can be Judge, Jury, and Executioner: Justice in the Age of Artificial Intelligence* (Singapore, World Scientific, 2021).

and automated decision-making in legal adjudication – their deployment in the hope of advancing adjudicative ideals such as the rule of law is ‘nothing more than wishful thinking’.<sup>28</sup> In addition, the risk of dysfunction is compounded by the possibility that an AI system operating in real-world scenarios might encounter unforeseen obstacles that impair its effective operation. Among them is the risk that litigants may be able to ‘game the system’, using strategies such as the selective suppression of information, or the making of false or irrelevant factual claims, in order to skew the operation of an AI adjudicative tool unfairly in their favour.

A second kind of worry is that even if AI adjudicative tools are reliably capable of yielding legally sound decisions as outputs, their widespread use may have various undesirable systemic side-effects, where the undesirability is broadly related to rule of law values. Concerns under this heading include the following: (a) the potential attrition and loss of human capacities in the domain of legal adjudication that would result from the ever-increasing deployment of AI tools. The idea here is that deliberation that leads to the making of an actual legal decision furnishes a setting for the cultivation and exercise of one’s rational powers in legal adjudication in a manner that is not replicated by hypothetical discussion of the very same subject-matter. The upshot is not only that humans are deprived of a domain for the pursuit of communal engagement and individual excellence, but also the knock-on effect of diminishing our capacity to subject AI adjudicatory tools to effective critical scrutiny, including with respect to their compliance with the rule of law, thereby converting them into an instrument of technocratic regulation over which there is limited democratic control; (b) the eventual dominance of AI tools in adjudication may lead to the de facto privileging of certain styles of legal reasoning which, albeit consistent with the rule of law, are not the only styles that are so consistent. The worry here is that human creativity and choice in the elaboration of a rule of law compliant judicial culture is unduly constricted;<sup>29</sup> and (c) the increasing prominence and prestige of AI-based adjudication may lead to a sense of disillusionment and alienation on the part of human beings with respect to law as a domain in which they have a valuable role to play, including in upholding the rule of law.<sup>30</sup>

In addition to these rule of law focussed ‘output-based’ reservations, I have already referred to ‘output-based’ concerns that are not strictly encompassed within the rule of law, such as equity and mercy. But even if we register the

<sup>28</sup> R Reich, M Sahami and JM Weinstein, *System Error: Where Big Tech Went Wrong and How We Can Reboot* (London, Hodder & Stoughton, 2021) 102.

<sup>29</sup> For the potential impact of AI in skewing the culture of legal adjudication, see Re and Solow-Niderman (n 8) 240. Much here turns on the extent to which judges, lawyers, legal academics and others will find themselves deferring to AI-style deliberation. For some scepticism about ‘automation bias’ in the case of judicial sentencing, see J Zerilli, ‘Algorithmic Sentencing: Drawing lessons from Human Factors research’, in J Ryberg, J Roberts and J de Keijser (eds), *Principled Sentencing and Artificial Intelligence* (Oxford, Oxford University Press, 2020).

<sup>30</sup> Re and Solow-Niderman (n 8) 240.

force of these objections, we may feel a lingering worry that ‘outputs’ – whether thought of as decisions yielded by AI adjudicative tools, or the consequences of deploying such tools – are not the only matters in play in seeking to realise the rule of law.<sup>31</sup> After all, we often care not only about the decisions that are outputs or their causal consequences, but *how* these decisions are reached. This is the idea I will explore in the rest of this chapter in relation to the desideratum of congruence between official decision and applicable law.

## V. PROCEDURE MATTERS

We do not only have first-order reasons to do certain things or achieve certain outcomes, we also have second-order reasons to do or achieve them in certain ways rather than others. In choosing a life partner, we have reason not only to make a good decision, but also reason to reach *our own* decision rather than bowing to the dictates of a benevolent parent or even a benevolent government data analyst – and this is so even if we reasonably believe those other people are endowed with superior judgment than us on such matters. How these first- and second-order reasons are to be balanced against each other, in cases in which they pull in different directions, is a further question. Still, at this point we can ask: does the rule of law requirement of congruence between rule and decision bear not only on *whether* decisions conform with the law but also on *how* they come to do so? Now, a hard-core consequentialist view denies any inherent significance to such procedural concerns. This is in line with Volokh’s injunction that ‘what matters is the result, not the process’. The implication is not that the value of outputs systematically trumps that of procedures but that the latter have no inherent value; they simply *do not matter* over and above the question of whether they lead instrumentally to a good decision or other outcome. I am going to set aside this scorched earth approach as obviously incorrect.

Sunstein and his co-authors, by contrast, seem to take a more nuanced view. In a chapter entitled ‘Dignity’ they discuss sources of resistance to algorithmic decision-making. One of these is a conception of due process which ‘might seem to require an opportunity for a face-to-face interaction in which a human being, authorized to exercise discretion, considers a wide range of factors [ie beyond following rules set down]’.<sup>32</sup> This ‘dignity’ concern can be elaborated from two

<sup>31</sup>The tendency to evaluate AI exclusively, or primarily, by reference to its outputs is not unique to the legal context: ‘To a large extent, AI is judged by the utility of its results, not the process used to reach those results. This signals a shift in priorities from earlier eras, when each step in a mental or mechanical process was either experienced by a human being (a thought, a conversation, an administrative process) or could be paused, inspected, and repeated by human beings.’ HA Kissinger, E Schmidt and D Huttenlocher, *The Age of AI and Our Human Future* (London, John Murray, 2021) 107.

<sup>32</sup>Kahneman, Sibony and Sunstein (n 22) 340.

perspectives. The person subject to decision-making seeks the satisfaction of a face-to-face interaction in which their specific circumstances can be taken into account, with the possibility of the decision-maker exercising discretion to depart from the rigid application of rules on grounds such as mercy. From the decision-maker's perspective, there is the sense of dignity that comes from having the capacity to exercise situation-specific judgment in departing from fixed rules. Elaborating on this latter perspective, Sunstein and his co-authors write:

In general, people in positions of authority do not like to have their discretion taken away. They may feel diminished as well as constrained – even humiliated. When steps are taken to reduce their discretion, many people will rebel. They value the opportunity to exercise judgment; they might even cherish it. If their discretion is removed so that they will do what everyone else does, they might feel like cogs in a machine ... Their jobs look more mechanical, even robotic. Who wants to work in a place that squelches your own capacity to make independent decisions?<sup>33</sup>

Although Sunstein and his co-authors seek to parry the principled force of these dignitarian procedural concerns, they concede the relevance of these concerns as sources of psychological resistance to algorithmic decision-making at the level of implementation.<sup>34</sup> This is why their recommendations largely consist in introducing algorithms as an aid to human decision-making, rather than the wholesale replacement of human decision by algorithms.

However, there are two general difficulties that arise for the approach of Sunstein and his co-authors. First, they are wrong to imply that giving leeway to human decision-makers to depart from rules is inherently noisy. Instead, whether noise ensues will depend on whether there are objective reasons for departing from rules in particular cases and whether human decision-makers reliably track those reasons. The whole point of the equity tradition is that deviation from legal rules in order to avoid their yielding an injustice in unanticipated circumstances does not lead to decisions unconstrained by objective reasons, but rather is itself a demand of reason. (The authors seem to make the same error about mercy, in addition to the common error that mercy is necessarily and extra-legal consideration, whereas it is obvious that legal rules can themselves embody considerations of mercy, for example, sentencing guidelines that treat an offender's repentance as a ground for leniency).

But the second and more important point is that there are process considerations at play even when we restrict our sights to the following of rules, rather than the exercise of human discretion to depart from rules. Moreover, these procedural considerations might play out differently depending on whether the rules are being followed by a human being versus an AI adjudicative tool. In other words, Sunstein and his co-authors are mistaken to assume that

<sup>33</sup> *ibid* 347.

<sup>34</sup> *ibid* 371.

discretion to depart from rules is the exclusive focus for the kind of dignitarian, process focussed-concern that is prompted by AI-based adjudication. If so, reflection on AI adjudicatory tools can highlight an aspect of the rule of law that is typically obscured from view, precisely because non-human legal adjudication was not a realistic prospect in the past.<sup>35</sup> If there is such a value, or set of values, and if humans decision-makers have a (greater) capability to realise them, then this would not necessarily definitively settle the question whether or not, in a given context, automated decision-making should be permitted. But it would be an important, albeit commonly neglected, consideration that is relevant to answering that question.

## VI. THE VALUE OF THE RULE OF LAW AND AI DECISION-MAKING

To see how AI-based decision-making might inhibit the full realisation of the rule of law, we must go beyond identifying the desiderata that comprise the rule of law's content in order to interpret them in light of the underlying values that they serve. Now, I believe that the formal and procedural desiderata comprehended within the thin account of the rule of law are primarily in the service of the value of rational autonomy.<sup>36</sup> The thin conception is not merely a shapeless list of valuable features of a legal system; instead, its desiderata are unified in two ways: they are all procedural-formal in character, and they serve the core value of respect for the rational autonomy of those subject to the law. Of course, there are other values – such as predictability, liberty, and democracy – that compliance with the rule of law also serves. But my contention is that respect for rational autonomy is the core value that such compliance realises, and that these other values are secondary, and often parasitic, in relation to it.

The rational autonomy in question is the capacity of human beings to step back in thought from their desires and inclinations, or from the influence of established patterns of behaviour, peer pressure, and so on, in order to assess the pros and cons of any given course of action in light of the objectively salient reasons, to reach an all-things-considered judgment as to the right course of action in light of that assessment, and to decide to follow that judgment in a particular case.<sup>37</sup> The thin account of the rule of law demands respect for

<sup>35</sup>For an excellent discussion of procedural, and not merely formal, aspects of the rule of law, see J Waldron, 'The Rule of Law and the Importance of Procedure', in J Fleming (ed), *Getting to the Rule of Law*, NOMOS vol 50 (New York, New York University Press, 2011). Waldron discusses procedural dimensions of the rule of law beyond Fuller's eight desiderata, whereas I focus below on procedural features of the congruence desideratum.

<sup>36</sup>Tasioulas (n 13) 121.

<sup>37</sup>There are various ways of elaborating on this idea of rational autonomy, but two elaborations that I find very attractive are to be found in J Raz, *The Morality of Freedom* (Oxford, Oxford University Press, 1986) chs 14–15 and J Griffin, *On Human Rights* (Oxford, Oxford University Press, 2008) ch 8. Both accounts stress that autonomy is an ingredient of human well-being and that it involves

rational autonomy by imposing formal-procedural requirements on how law plays into the practical deliberation of human beings. Compliance with the rule of law means that those subject to the law are able to factor the law that applies to them – irrespective of the merits of its specific content – into their exercises of rational autonomy because this law satisfies the desiderata of clarity, stability, non-contradictoriness, etc. But, of course, the satisfaction of such requirements will mean little if, having adopted a course of action partly on the basis of a grasp of what the law is, their legitimate expectations are frustrated by legal officials who do not subsequently go on to apply the law that had been laid down. Hence the requirement of congruence.

It is in these terms that we should understand Fuller’s observation that deviations from the rule of law are ‘an affront to man’s dignity as a responsible agent’,<sup>38</sup> since the capacity for rational autonomy is a vital dimension of the dignity, or intrinsic moral worth, of human beings. Of course, full respect for the rational autonomy of human beings goes far beyond compliance with the rule of law. For example, it additionally requires that they be protected from such evils as torture, slavery, and poverty, and that they be afforded adequate access to such goods as education, health care, and work. But the rule of law, on the thin conception, is a distinctive, formal-procedural kind of way in which law can respect the rational autonomy of its subjects.

Now, one general formulation of the procedural concern about the ability of AI adjudicative systems to fulfil the rule of law’s requirement of congruence is to ask the following question: does it take a legal decision-maker with the capacity for rational autonomy to respect fully the rational autonomy of the human beings that are subject to their decisions? But before pursuing this concern, it should be clear that the kind of adjudicative AI tools that I have in mind lack the rational autonomy that I have attributed to human beings. This is admitted even by those, such as Stuart Russell, who take seriously the prospect of technology based on Artificial General Intelligence (AGI) emerging in the not-too-distant future. The notion of ‘intelligence’ Russell deploys remains an instrumental one, in which the autonomous setting of valuable goals, and the autonomous identification of moral constraints on the pursuit of a given goal, is treated as extraneous to the operations of intelligence. Hence, maximising the number of paper clips in existence could be the ultimate goal programmed into an AGI system. This is why the potential emergence of such systems is so disturbing, given that their form of intelligence would not be autonomously responsive to reasons for not achieving their goals in ways that are catastrophic for human interests.<sup>39</sup> This contrasts with the notion of rational autonomy

choice from a worthwhile set of options in light of an appreciation of their objective pros and cons. The former feature, in particular, sets them apart from more austere Kantian approaches.

<sup>38</sup> Fuller (n 11) 162.

<sup>39</sup> S Russell, *Human Compatible: AI and the Problem of Control* (London, Penguin, 2020) 167.

articulated above, which encompasses not just the evaluation of instrumental means to given ends, but also the evaluation of candidate ends and means in terms of objective ethical values.<sup>40</sup>

In what follows, I argue that the use of AI tools in place of human judges undermines, in three ways, the kind of respect for rational autonomy that the desiderata of the rule of law are intended to secure. Specifically, it undermines the explainability, the accountability, and the reciprocity or community that is intimately bound up with the ideal fulfilment of the rule of law, and – in particular – the desideratum of congruence between law and official decision. Some of the concerns that I will articulate are already implicit in the passage from Plato's *Laws* that serves as the epigraph of this chapter. The free doctor referred to by Plato practices medicine in light of a general scientific understanding of human health which enables him to explain to his patient why he prescribes the treatment that he does, thereby enabling the patient to grasp the need to undergo the prescribed treatment and, in a real sense, to become a participant in their own cure rather than merely a passive beneficiary of it. The patient has at least the opportunity to enter into a dialogue with his doctor – a dialogue from which the doctor also potentially stands to be enlightened. The relationship of free doctor to free patient is, in the ideal case, a form of 'partnership' premised on the joint exercise of their rational capacities in advancing a shared commitment to improving the patient's health, which is itself a rationally chosen goal.

The slave doctor, by contrast, acquires his skill through 'trial and error' and mimicking the prescriptions of the free doctor – something that resembles the operations of a machine learning system trained on a vast store of data. The slave doctor expects his prescriptions to be accepted on the basis of authority ('with all the arrogance of a tyrant'), without offering any reasoned explanation of them or making any attempt to engage the patient's understanding or to secure their informed consent. Plato deploys this medical analogy to advance the case for introducing preambles (*prooimia*) that explain the point of legal enactments. This proposal conjoins command with persuasion in the act of legislation ('the dual approach'), in a manner that mirrors the free-born doctor's fusion of prescription and persuasion.<sup>41</sup> But this medical analogy is relevant to

<sup>40</sup> Russell seems committed to the denial of objective ethical values on the basis that the existence of such presupposes the possibility of inter-species agreement on them: 'it presupposes that there is a "right" objective out there in the world; it would have to be an objective on which iron-eating bacteria and humans and all other species agree, which is hard to imagine'. Russell (n 39) 168. But this requirement of inter-species agreement seems dubious at best, which then raises the question why a genuinely 'super-intelligent' automated system would not register the normative force of objective values, and hence not be liable to pose the kind of existential risk to humans that Russell posits. For a different characterisation of ethical objectivity, in terms of the availability of 'vindicatory explanations' of coming to believe a given ethical proposition, see D Wiggins, *Ethics: Twelve Lectures on the Philosophy of Morality* (Cambridge MA, Harvard University Press, 2006) chs 11–12.

<sup>41</sup> For a helpful discussion of this medical analogy in *The Laws*, see J Annas, *Virtue and Law in Plato and Beyond* (Oxford, Oxford University Press, 2017) 91–105.

the process of legal adjudication, as well as that of law-making. Although slave doctors operate on the basis of human rather than artificial intelligence, the shortcomings in his way of ministering to their patients highlighted by Plato offer helpful sign-posts as to the kind of procedural shortcomings that afflict AI adjudicative tools. Whereas in the former case these shortcomings are traceable to the failure of a human being both to exercise and respect rational autonomy to a sufficient degree, in the latter case they stem from the inherent inability of AI adjudicative tools to do so.

*Explainability.* Central to what we might think of as the paradigm case of legal adjudication, one in which the values integral to it are most fully realised, is the giving of reasons for decisions in particular cases by the adjudicator. This involves appealing to relevant considerations, including crucially the applicable law, that justify the adjudicator's decision. The applicable law comprises considerations that the litigants in the case were able to grasp and factor into their rational deliberation in advance of their own decision-making.<sup>42</sup> This means that, for the rule of law to be fully realised, the decisions made by the officials need to be justifiable to those subject to them, and justifiable (among other respects) in terms of the sound application of existing law. In this way, the meeting of the congruence requirement is not just a matter of bare conformity of the decisions with law, but of the processes through which those decisions are made. The provision of a justification not only helps the subject of the law grasp the meaning of the official decision, it also puts them in a better position to assess that decision as law-compliant and to challenge it if it is not. Moreover, especially in an adjudicatory context, the proffered justification can provide them with some assurance that their arguments have been taken into account and engaged with, whether or not they were ultimately successful in the case at hand. It also gives them assurance that the decision was not merely coincidentally congruent with the law, but instead robustly grounded in the law because it was arrived at *precisely because* it is in accord with the law. In these various ways, the giving of legal reasons for a decision, in order to demonstrate the satisfaction of the congruence requirement, honours the rational capacities of those subject to those decisions.

Now, what I have said here about the importance of giving reasons for decisions in legal adjudication arises as an issue in relation to automated decision-making most often under the heading of 'explainability'. So let me pursue some worries about AI-powered adjudicatory tools under this heading in four steps, focussing on the deep-learning style of machine learning that has achieved prominence in recent years.

The first worry is whether there is an explanation available for decisions reached by the AI system. Now, such an explanation may be known to the

<sup>42</sup> Many of the points in this paragraph are to be found in L Fuller, 'The Forms and Limits of Adjudication' (1978) 92 *Harvard Law Review* 353, 366–67 and 388.

technicians who devised or operate the system. But it may not be accessible to others owing to restrictions imposed by intellectual property rights in the algorithm or to concerns about security (eg, that publicising the algorithm will lead to gaming of the system by malevolent actors). Alternatively, it may be that no adequate explanation is available even to those who created the AI system, since they cannot fully grasp why it reaches the decisions that it does as it revises its algorithm over time in response to new data and feedback loops. So, even if an AI-based decision happens to be congruent with the existing law, there may be no available explanation of how it was generated. Indeed, as Emily Berman has pointed out, ‘the more complex and powerful an algorithm, the more opaque it is likely to be’,<sup>43</sup> with the result that we face a trade-off between explainability and correctness of decisional outputs. The second worry is that even if an explanation of the decision is available, it may be of such technical complexity as to be epistemically inaccessible to ordinary citizens not versed in the science of machine learning. There is no real comfort in knowing that an explanation exists, but having no clue as to its nature.

Third, even if an explanation exists and is accessible to a minimally adequate degree, there is still a further question as to whether it is an explanation of the *right kind*, in the sense of being one that *justifies* the decision that has been made. Machine learning processes may reach the same results as ordinary human legal reasoning, but the means through which they do so differ radically. Machine learning AI systems deploy a statistical process of pattern-detection, one that discerns statistical correlations in a vast amount of data, and on this basis make classifications or predictions relating to new cases. For example, a deep learning system that was trained to distinguish pictures of wolves from pictures of huskies was discovered to be using the presence of snow in a picture as a determining factor – it may have yielded overwhelmingly correct classifications, but for the wrong reasons.<sup>44</sup> This is a categorically different process from the essentially normative enterprise of *justifying* a decision in a particular case by reference to the relevant reasons for that decision.<sup>45</sup> Now, compare in this respect the automated system Lex Machina, which is apparently able to predict the probability of success in US patent litigation more accurately than patent lawyers. Its predictions are not based on past judgments in law reports, but rather on ‘data *about* over 100,000 past cases – features such as the names of

<sup>43</sup>E Berman, ‘A Government of Laws and Not of Machines’ (2018) 98 *Boston University Law Review* 1, 277 and 315.

<sup>44</sup>M Ribeiro, S Singh and C Guestrin, ‘“Why Should I trust You?” Explaining the Predictions of Any Classifier’ (Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining, 2016) 1135–44, discussed in J Zerilli, ‘Explaining Machine Learning Decisions’ (2021) 89 *Philosophy of Science* 1.

<sup>45</sup>Hence the very limited consolation to be derived from claims by writers such as Sunstein, that algorithms are in certain respects less of a ‘black box’ than ordinary human reasoning processes. The matter is not simply about the existence and accessibility of an explanation, but of its intrinsic nature.

the judges, the law firms and the lawyers, the nature and value of the claims, and so forth'.<sup>46</sup> Now, it may well be that such a system has great value as a predictor of the outcome of litigation, and could even be conceivably useful as a decider of cases. But the explanation of its decision-making obviously does not consist in the application of existing law to the case at hand.

Finally, even if an AI adjudicatory tool can generate a normative justification for its decision *ex post*, there is still the question of whether sensitivity to that justification was *causally efficacious* in bringing about the decision. As we have seen, Volokh tries to side-step this issue by saying, in relation to an imagined AI tool that passes his adjudicative Turning test, that 'I have carefully tried to avoid saying that the AI judge should explain why *it* reached the result it reached, saying instead that the AI judge should articulate reasons *supporting* the decision.'<sup>47</sup> This caveat is readily understandable, given the statistical nature of ML decision-making processes. But it is one thing to uphold the requirement of congruence by showing that the sound application of the law to the fact situation played a causal role in the judge's decision. It is quite a different thing, a fool's gold version perhaps, to be given an *ex post* rationalisation of the decision that is causally inert, when the real cause of the decision is quite different.<sup>48</sup> This point is all the more stark in light of the fact that in some cases the AI system that generates the *ex post* explanation is entirely distinct from the one that made the original decision.

*Answerability.* Supposing the challenges around explainability could somehow be addressed or suitably mitigated – perhaps by the use of what is colloquially known as 'Good Old Fashioned AI', which involves expert systems whose rules of operation can be stated in natural language – there is a further issue that applies to AI adjudicative tools generally. Even if a legal decision rendered by an AI adjudicative tool is congruent with the law, and generated through a process whereby the legal justification for it is causally efficacious, there is yet another difference between the human judge and the AI tool. The core value underlying the rule of law, as we have seen, is respect for rational autonomy in the exercise of governance through law. But this implicates not just the rational autonomy of ordinary citizens who are subject to the law, but also that of officials, including judges. So, in the case of a human judge, we have the idea of the exercise of rational autonomy on at least three levels: (a) in the commitment to follow the law as a set of standards that give both the judge and the litigants reasons for compliance; (b) in the commitment to congruence

<sup>46</sup> Susskind (n 2) 282.

<sup>47</sup> Volokh (n 16) 164.

<sup>48</sup> From an *ex ante* perspective there is the related point that a valuable and distinctive kind of predictability – one achieved not by merely grasping an empirical regularity in official behaviour but by grasping the normative considerations that are causally effective in generating official decisions – is secured when officials are actually guided by existing law, see J Tasioulas, 'The Legal Relevance of Ethical Objectivity' (2002) 47 *American Journal of Jurisprudence* 211, 244–54.

between law and decision, both as a general matter, and in specific instances of adjudication; and (c) in the commitment to a certain decision in a specific case as being in accordance with the applicable law. By having the capacity to make these commitments through the exercise of their rational autonomy, human judges can take responsibility for their decisions and, in consequence, can be regarded as answerable for those decisions by litigants and the broader community in whose name judges officially act.

Being subject to a decision by a decision-maker who is answerable for it, in this way, has intrinsic value. This is indicated by the fact that we believe this value is present even when we disagree – even vehemently so – with the decision that has been rendered.<sup>49</sup> We can, for example, imagine a disappointed litigant nonetheless valuing the fact that the judge in their case acted in light of their appreciation of the applicable law and took responsibility for their decision. This is a significant element of the *procedural* value of legal adjudication that is relatively autonomous from any concern with its output value, part of the reason why litigants may seek their ‘day in court’ even if they believe their prospects of receiving a favourable judgment are dim – they seek not just the binding legal resolution of their case, but decision through a process in which an official, as the mouthpiece of the legal order, is properly answerable for it. By contrast, an AI adjudicative tool cannot be regarded as answerable in the same way. This is because the idea of making a commitment is inapplicable in its case. It cannot stand back from an array of options, such as whether to commit morally to the legal system, or to the requirement of congruence, on the basis of deliberation about the pros and cons of doing these things. In other words, doing so is to display a kind of rational autonomy, a notion central to contemporary characterisations of human dignity, that is not applicable in the case of AI technology as we currently know it, or even of the AGI-type anticipated by Russell and others.

The psychologist Mandeep Dhimi reports of the criminal offenders she has worked with: ‘Even knowing that the human judge might make more errors, the offenders still prefer a human to an algorithm. They want that human touch.’<sup>50</sup> This way of putting it can make it sound as though the offenders are in the grip of nostalgic sentimentalism, and Dhimi stresses that she takes an opposite view to these offenders, preferring algorithmic sentencing to that provided by a fallible human being. But there is, I think, more than a grain of wisdom in

<sup>49</sup> Contrast, here, the approach of Volokh (n 16) 163, which takes the idea of answerability to be merely instrumentally valuable as a means of generating sound decisional outputs: ‘The human judge’s oath, the human judge’s sense of individual responsibility, the President’s or Senate’s or voters’ evaluation of the judge’s integrity – those are all valuable, but they are means to an end, the end being results that we think are fair, sound, and efficiently produced. If AI judges’ opinions persuade us of their fairness and soundness more than human judges’ opinions do, and at the same time come with less cost and delay, then that should be sufficient.’

<sup>50</sup> Quoted in H Fry, *Hello World: How to be Human in the Age of the Machine* (London, Doubleday, 2018) 76.

the offenders' craving for a human judge, despite human fallibility with respect to decisional outputs. And one aspect of this wisdom is the feeling of being respected as a rational agent in having one's conduct and subsequent argument judged by a fellow human being who can take responsibility for their decision and be held answerable for it. So, Sunstein and his co-authors are right to observe that human decision-makers often regard algorithmic decision-making as 'dehumanizing and an abdication of their responsibility'.<sup>51</sup> But rather than this being a brute psychological tendency with dubious normative credentials, as the authors seem to suggest,<sup>52</sup> we should take it seriously as the intimation of a genuine intrinsic value. Of course, how much weight this value should be accorded, and the extent to which it can be outweighed by gains in the output-related aspects of congruence, are further questions.

Now, of course, the advocates of AI adjudicative systems can respond to the point about answerability by arguing that, ultimately, there is always human answerability because some human agent or group of agents will have taken the decision to deploy a specific AI system in a defined decision-making domain. Moreover, behind the decision to deploy a given AI adjudicative system, and standing in complex relations with it, will be a series of consequential human choices about the design of the system – selection of training data to ensure completeness and accuracy, choice of model and among different kinds of algorithms, choices regarding how to interpret a system's outputs or assess its performance. Human responsibility, therefore, is hardly expunged under their proposed scheme. This is a vitally important point. Still, the distal answerability for the design and implementation of a system of automated decision-making differs from the actual human judge in the particular case being answerable for the specific decision in that very case. What is at issue here is not so much a 'responsibility gap', per se, but the absence of a certain *kind* of responsibility that we have reason to value. It is one thing to take responsibility for the design and general adoption of an automated system that yields often unanticipated decisions in particular cases, it is quite different actually to make a decision in such a case and stand behind it. In the former case, there is a diffuse answerability to all those impacted by the automated system as a group; in the latter case, the answerability to the person affected by the specific decision is more direct and personal.

*Reciprocity.* Let me proceed to a third value secured by the rule of law, in virtue of its grounding in a certain kind of respect for rational autonomy. This is the value of *reciprocity*, by which I mean not an exchange that is in some sense in the mutual self-interest of the parties to it, but rather the mutual recognition and exercise of rational agency secured by adherence to the rule of law.

<sup>51</sup> Kahneman, Sibony and Sunstein (n 22) 134.

<sup>52</sup> 'The goal of judgment is accuracy, not individual expression ... judgment is not the place to express your individuality' (emphasis added). Kahneman, Sibony and Sunstein (n 22) 371.

It finds an echo in the idea of ‘partnership’ and ‘dialogue’ that characterises Plato’s account of the relationship between free citizen doctors and patients.

As theorists like Fuller and Finnis have argued, when officials act on the basis of their commitment to the application of legal standards, and ordinary citizens are able to anticipate how officials are liable to impinge upon their activities by means of a corresponding grasp of those same standards, a valuable form of reciprocity is made possible. The exercise of the capacity for rational autonomy possessed by officials dovetails with the operation of the same capacity on the part of citizens. The exercise by officials of their rational autonomy in committing to the application of the law in line with the desideratum of congruence finds its counterpart in the exercise of rational autonomy by citizens in guiding their conduct on the basis of a grasp of those legal standards that officials will apply. The answerability of officials to the law is paralleled by the similar answerability of ordinary citizens. Law in this way can form a plateau on which a valuable form of the mutual recognition and exercise of rational autonomy plays out.

This reciprocity is vividly manifested in the context of a legal trial, which is a form of institutionalised ‘dialogue’ of the kind envisioned by Plato in the idea of ‘partnership’ between doctor and patient, although here the subject matter is right and wrong conduct rather than health. The parties before the judge are given the opportunity to make their case as to why, among other things, their conduct adhered to the laws laid down; meanwhile, the judge, in giving his reasons for his decision, has the opportunity to show that he has taken those arguments into account in exercising his own rational capacities in determining what the law requires in the given case.<sup>53</sup>

This reciprocity can be understood as operating at various levels of specificity, but two are especially salient. The first is at the cosmopolitan level of human beings exercising their rational capacities in seeking to align their conduct – whether as officials or citizens – with genuine reasons for action, including those generated by the law. The second is at the political level of members of a shared political community exercising their rational capacities in the same way in relation to the law of that community, which in the paradigm case is a concretisation of genuine demands of reason for a specific time, place and community. The institution of the jury can be helpfully seen here as an attempt to widen the reciprocity by allowing ordinary citizens who are not legally expert to play an adjudicative role. Now, it might be thought that, by a further natural progression, this idea ought to extend to reciprocity among adult members of the community in coming together to *enact* their own laws by means of collective deliberation as free and equal citizens, which is the core idea of democracy.

<sup>53</sup>For a development of the idea of the criminal trial as having value as a *process* of calling offenders to account, not just as an instrumental means of determining guilt or innocence, see RA Duff and others, *The Trial on Trial: Towards a Normative Theory of the Criminal Trial*, vol 3 (Oxford, Hart Publishing, 2007).

And, unlike Plato, who in *The Republic* lauded rule by an expert philosopher class and in *The Laws* recommended subjection to an unmodifiable code of laws, I believe that it does extend in this way. But I also think this step takes us beyond the rule of law proper, while also revealing that democracy and the rule of law nonetheless share a grounding value in respect for rational autonomy.

Nothing like this valuable form of reciprocity, whereby the exercise of rational autonomy in being guided by the law unites officials and citizens alike, can be found when a judge lacks those characteristic human capacities. This is one form of the familiar apprehension about the alienated nature of life in a ‘dehumanised’ world shaped by automated decision-making. Moreover, it is reasonable to suppose that reciprocity has intrinsic value – as does answerability – that is to a large degree independent of the success of citizens or officials in reaching decisions (‘outputs’) that actually conform with the law. It is an essentially procedural consideration, though obviously one that can have its value diminished or perhaps even completely obliterated in certain circumstances, for example, if the decisions reached are systematically out of kilter with the law or the content of the law is grossly unjust.

## VII. CONCLUSION

There are various general responses that the proponent of AI in legal adjudication might make to the argument in this chapter. One is the claim that I have unfairly rigged the dialectic by comparing potential AI adjudicative tools with idealised rather than actual human judges. In this vein, it will be pointed out that the workings of the human judicial psyche themselves often approximate to a ‘black box’; that judicial explanations of decisions are often little better than ex post rationalisations; and that formidable personal and structural obstacles hinder the realisation of meaningful forms of answerability and reciprocity.<sup>54</sup> Of course, all this is a matter for investigation. But as Meghan O’Gieblyn has written, we should be wary of rushing to embrace the conclusion that it is ‘just as reasonable to trust the opaque logic of an algorithm as it is to trust our own minds’,<sup>55</sup> not least because inter-personal interpretation benefits from the hermeneutical bridge forged by a shared human nature.<sup>56</sup> Moreover, whatever the truth of the matter about human beings falling short of the aspirations of

<sup>54</sup> See, eg. on opacity, AZ Huq, ‘A Right to a Human Decision’ (2020) 106 *Virginia Law Review* 611, 645.

<sup>55</sup> M O’Gieblyn, *God Human Animal Machine: Technology, Metaphor, and the Search for Meaning* (New York, Doubleday, 2021) 210.

<sup>56</sup> On the hermeneutical significance of a shared human nature, see D Wiggins, *Continuants – Their Activity, Their Being, and Their Identity. Twelve Essays* (Oxford, Oxford University Press, 2016) 91: ‘[O]ur sharing in a given specific animal nature and a law-sustained mode of activity is integral to the close attunement of person to person in language and integral to the human sensibilities that make interpretation possible’.

explainability, answerability, and reciprocity, there is value in human legal adjudication simply in virtue of the fact that inherent within it is the *possibility* of realising these ideals. This enables it to be the focus of a valuable form of hope that is out of place in the case of AI adjudicative systems.

A second response is that nothing in my argument establishes a general, all-things-considered case against the use of AI adjudicative systems in place of human judges. This is because the procedural limitations of such systems that I have identified need to be set against their potential benefits, such as enhanced consistency with the rule of law in terms of decisional outputs, as well as the widely-touted efficiency gains. This response is correct, so far as it goes. But my main aim was to show that the case for AI adjudicative systems has serious limitations even with respect to the most promising value that could be appealed to in its defence, that is, the rule of law. Still, the weighing up of pros and cons envisaged by the second response will plausibly need to be done in a piecemeal fashion, focussed on specific proposals to deploy AI adjudicative tools in specific contexts of decision-making. There is no reason to suppose that the all-things-considered verdict on the deployment of AI adjudicative tools will be totalising in character, endorsing either their wholesale adoption or rejection.

Elaborating on this piecemeal and contextual approach, a third response is to retreat from the original proposal, which was that of deploying AI adjudicative systems as a wholesale alternative to human judges. Instead, the revised proposal would countenance a variety of ways in which AI adjudicative systems might be deployed. Sometimes as replacements for human judges, sometimes in one of the various modalities of working in conjunction with them.<sup>57</sup> Whether the deployment of AI adjudicative tools is acceptable, and what form it should take, will turn on factors such as the domain of law in question, the values at stake, availability of resources, and wider issues of institutional design and cultural setting. This fall-back position is definitely worth exploring, but it enters into complex practicalities that are well downstream of the more abstract and foundational questions addressed in this chapter.

<sup>57</sup> See, eg, the discussion of human decision-makers being ‘out of the loop’, ‘over the loop’, and ‘in the loop’, in S Chesterman, *We, The Robots? Regulating Artificial Intelligence and the Limits of Law* (Cambridge, Cambridge University Press, 2021) 87.

