11

What's in the Box?

The Legal Requirement of Explainability in Computationally Aided Decision-Making in Public Administration

Henrik Palmer Olsen,[*] *Jacob Livingston Slosser,*[**] *and Thomas Troels Hildebrandt*[†]

11.1 INTRODUCTION

As the quality of AI[1] improves, it is increasingly applied to support decision-making processes, including in public administration.[2] This has many potential advantages: faster response time, better cost-effectiveness, more consistency across decisions, and so forth. At the same time, implementing AI in public administration also raises a number of concerns: bias in the decision-making process, lack of

[*] Associate Dean for Research, Professor of Jurisprudence, iCourts (Danish National Research Foundation's Centre of Excellence for International Courts) at the University of Copenhagen, Faculty of Law; henrik.palmer.olsen@jur.ku.dk. This work was produced in part with the support of Independent Research Fund Denmark project PACTA: Public Administration and Computational Transparency in Algorithms, grant number: 8091–00025

[**] Carlsberg Postdoctoral Fellow, iCourts (Danish National Research Foundation's Centre of Excellence for International Courts) at the University of Copenhagen, Faculty of Law; jacob.slosser@jur.ku.dk. This work was produced in part with the support of the Carlsberg Foundation Postdoctoral Fellowship in Denmark project COLLAGE: Code, Law and Language, grant number: CF18-0481.

[†] Professor of Computer Science, Software, Data, People & Society Research Section, Department of Computer Science (DIKU), University of Copenhagen; hilde@di.ku.dk. This work was produced in part with the support of Independent Research Fund Denmark project PACTA: Public Administration and Computational Transparency in Algorithms, grant number: 8091–00025 and the Innovation Fund Denmark project EcoKnow.org.

[1] AI is here used in the broad sense, which includes both expert systems and machine learning as well as hybrid models. Various webpages contain information about how AI and Machine Learning may be understood. For an example, see www.geeksforgeeks.org/difference-between-machine-learning-and-artificial-intelligence/.

[2] See also Jennifer Cobbe, 'Administrative Law and the Machines of Government: Judicial Review of Automated Public-Sector Decision-Making' (2019) 39 *Legal Studies* 636; Monika Zalnieriute, Lyria Bennett Moses, and George Williams, 'The Rule of Law and Automation of Government Decision-Making' (2019) 82 *The Modern Law Review* 425. Zalnieriute et al. conduct four case studies from four different countries (Australia, China, Sweden, and United States), to illustrate different approaches and how such approaches differ in terms of impact on the rule of law.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

transparency, and elimination of human discretion, among others.[3] Sometimes, these concerns are raised to a level of abstraction that obscures the legal remedies that exist to curb those fears.[4] Such abstract concerns, when not coupled with concrete remedies, may lead to paralysis and thereby unduly delay the development of efficient systems because of an overly conservative approach to the implementation of ADM. This conservative approach may hinder the development of even safer systems that would come with wider and diverse adoption. The fears surrounding the adoption of ADM systems, while varied, can be broadly grouped into three categories: the argument of control, the argument of dignity, and the argument of contamination.[5]

The first fear is the loss of control over systems and processes and thus of a clear link to responsibility when decisions are taken.[6] In a discretionary system, someone must be held responsible for those decisions and be able to give reasons for them. There is a legitimate fear that a black box system used to produce a decision, even when used in coordination with a human counterpart or oversight, creates a system that lacks responsibility. This is the fear of the rubber stamp: that, even if a human is in the loop, the deference given to the machine is so much that it creates a vacancy of accountability for the decision.[7]

The second fear of ADM systems is that they may lead to a loss of human dignity.[8] If legal processes are replaced with algorithms, there is a fear that humans will be

[3] See, among various others, Virginia Eubanks, *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor* (St Martin's Press 2018); Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Broadway Books 2017).

[4] We find that some of the ethical guidelines for AI use, such as the European Commission's Ethics Guidelines for Trustworthy AI (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai) raise general concerns, but do not provide much guidance on how to address the concerns raised.

[5] These categories are generally sketched from Bygrave's analysis of the *travaux préparatoires* of Art 22 of the General Data Protection Regulation, which concerns explanation in automated processing and the Commission's reticence towards implementing fully automated systems exemplified in Art 15 of the Data Protection Directive. See the draft version at p 6–7 of the chapter on Art. 22: Lee A Bygrave, 'Article 22', 2019 *Draft Commentaries on 6 Articles of the GDPR (From Commentary on the EU General Data Protection Regulation)* (Oxford University Press 2020) https://works.bepress.com/christopher-kuner/2/download.

[6] A related but more legal technical problem in regards to the introduction of AI public administration is the question of *when* exactly a decision is made. Associated to this is also the problem of *delegation*. If a private IT developer designs a decision-system for a specific group of public decisions, does this mean that those decisions have been delegated from the public administration to the IT developer? Are future decisions *made* in the process of writing the code for the system? We shall not pursue these questions in this chapter, but instead proceed on the assumption that decisions are made when they are issued to the recipient.

[7] Elin Wihlborg, Hannu Larsson, and Karin Hedstrom, '"The Computer Says No!" – A Case Study on Automated Decision-Making in Public Authorities', *2016 49th Hawaii International Conference on System Sciences (HICSS)* (IEEE 2016) http://ieeexplore.ieee.org/document/7427547/.

[8] See e.g., Corinne Cath et al., 'Artificial Intelligence and the "Good Society": The US, EU, and UK Approach' [2017] *Science and Engineering Ethics* http://link.springer.com/10.1007/s11948-017-9901-7.

reduced to mere 'cogs in the machine'.[9] Rather than being in a relationship with other humans to which you can explain your situation, you will be reduced to a digital representation of a sum of data. Since machines cannot reproduce the whole context of the human and social world, but only represent specific limited data about a human (say age, marital status, residence, income, etc.), the machine cannot *understand* you. Removing this ability to understand and to communicate freely with another human and the autonomy which this represents can lead to alienation and a loss of human dignity.[10]

Third, there is the well-documented fear of 'bad' data being used to make decisions that are false and discriminatory.[11] This fear is related to the ideal that decision-making in public administration (among others) should be neutral, fair, and based on accurate and correct factual information.[12] If ADM is implemented in a flawed data environment, it could lead to systematic deficiencies such as false profiling or self-reinforcing feedback loops that accentuate irrelevant features that can lead to a significant breach of law (particularly equality law) if not just societal norms.[13]

While we accept that these fears are not unsubstantiated, they need not prevent existing legal remedies from being acknowledged and used. Legal remedies should be used rather than the more cursory reach towards general guidelines or grand and ambiguous ethical press releases, that are not binding, not likely to be followed, and do not provide much concrete guidance to help solve the real problems they hope to address. In order to gain the advantages of AI-supported decision-making,[14] these concerns must be met by indicating how AI can be implemented in public administration without undermining the qualities associated with contemporary administrative procedures. We contend that this can be done by focusing on how ADM can be introduced in such a way that it meets the requirement of explanation as set out in administrative law at the standard calibrated by what we expect legally out of human explanation.[15] In contradistinction to much recent literature, which focuses on the

[9] Meg Leta Jones, 'The Right to a Human in the Loop: Political Constructions of Computer Automation and Personhood' (2017) 47 *Social Studies of Science* 216.

[10] Karl M. Manheim and Lyric Kaplan, 'Artificial Intelligence: Risks to Privacy and Democracy' (Social Science Research Network 2018) *SSRN Scholarly Paper ID 3273016* https://papers.ssrn.com/abstract=3273016.

[11] For discussion of this issue in regards to AI supported law enforcement, see Rashida Richardson, Jason Schultz, and Kate Crawford, 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice' [2019] *New York University Law Review Online* 192.

[12] Finale Doshi-Velez et al., 'Accountability of AI Under the Law: The Role of Explanation' [2017] *arXiv:1711.01134 [cs, stat]* http://arxiv.org/abs/1711.01134.

[13] See, among others, Pauline T. Kim, 'Data-Driven Discrimination at Work' (2016) 58 *William & Mary Law Review* 857.

[14] See Zalnieriute, Moses, and Williams (n 2) 454.

[15] By *explanation*, we mean here that the administrative agency gives reasons that support its decision. In this chapter, we use the term *explanation* in this sense. This is different from *explainability*, as used in relation to the so-called 'black box problem'; see Cynthia Rudin, 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead' (2019) 1 *Nature*

right to an explanation solely under the GDPR,[16] we add and consider the more well-established traditions in administrative law. With a starting point in Danish law, we draw comparisons to other jurisdictions in Europe to show the common understanding in administrative law across these jurisdictions with regard to assuring administrative decisions are explained in terms of the legal reasoning on which the decision is based.

The chapter examines the explanation requirement by first outlining how the explanation should be understood as a *legal* explanation rather than a *causal* explanation (Section 11.2). We dismiss the idea that the legal requirement to explain an ADM-supported decision can be met by or necessarily implies mathematical transparency.[17] To illustrate our point about legal versus causal explanations, we use a scenario based on real-world casework.[18] We consider that our critique concerns mainly a small set of decisions that focus on legal decision-making: decisions that are based on written preparation and past case retrieval. These are areas where a large number of similar cases are dealt with and where previous decision-making practice plays an important role in the decision-making process (e.g., land use cases, consumer complaint cases, competition law cases, procurement complaint cases, applications for certain benefits, etc.). This scenario concerns an administrative decision regarding the Danish law on the requirement on municipalities to provide compensation for loss of earnings to a parent (we will refer to them as Parent A) who provides care to a child with a permanent reduced physical or mental functioning (in particular whether an illness would be considered 'serious, chronic or long-term'). The relevant legislative text reads:

> Persons maintaining a child under 18 in the home whose physical or mental function is substantially and permanently impaired, or who is suffering from serious, chronic or long-term illness [shall receive compensation]. Compensation shall be subject to the condition that the child is cared for at home as a necessary consequence of the impaired function, and that it is most expedient for the mother or father to care for the child.[19]

Machine Intelligence 206. As we explain later, we think the quest for black-box explainability (which we call *mathematical transparency*) should give way to an explanation in the public law sense (giving grounds for decisions). We take this to be in line with Rudin's call for interpretability in high-stakes decisions.

[16] See e.g., Sandra Wachter, Brent Mittelstadt, and Luciano Floridi, 'Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation' (2017) 7 *International Data Privacy Law* 76; Margot E. Kaminski, 'The Right to Explanation, Explained' (2019) 34 *Berkeley Tech. LJ* 189.

[17] See the debate regarding transparency outlined in Brent Daniel Mittelstadt et al., 'The Ethics of Algorithms: Mapping the Debate' (2016) 3(2) *Big Data & Society* 6–7.

[18] See the Ecoknow project: https://ecoknow.org/about/.

[19] § 42 (1) of the Danish Consolidation Act on Social Services, available at http://english.sm.dk/media/14900/consolidation-act-on-social-services.pdf. For a review of the legal practice based on this provision (in municipalities), see Ankestyrelsen, 'Ankestyrelsens Praksisundersøgelse Om Tabt Arbejdsfortjeneste Efter Servicelovens § 42 (National Board of Appeal's Study on Lost Earnings According to Section 42 of the Service Act)' (2017) https://ast.dk/publikationer/ankestyrelsens-praksisundersogelse-om-tabt-arbejdsfortjeneste-efter-servicelovens-ss-42.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

We will refer to the example of Parent A to explore explanation in its causal and legal senses throughout.

In Section 11.3, we look at what the explanation requirement means legally. We compare various national (Denmark, Germany, France, and the UK) and regional legal systems (EU law and the European Convention of Human Rights) to show the well-established, human standard of explanation. Given the wide range of legal approaches and the firm foundation of the duty to give reasons, we argue that the requirements attached to the existing standards of explanation are well-tested, adequate, and sufficient to protect the underlying values behind them. Moreover, the requirement enjoys democratic support in those jurisdictions where it is derived from enacted legislation. In our view, ADM can and should be held accountable under those existing legal standards and we consider it unnecessary to public administration if this standard were to be changed or supplemented by other standards or requirements for ADM and not across all decision makers, whether human or machine. ADM, in our view, should meet the same minimum explanation threshold that applies to human decision-making. Rather than introducing new requirements designed for ADM, a more dynamic communicative process aimed at citizen engagement with the algorithmic processes employed by the administrative agency in question will be, in our view, more suitable to protecting against the ills of using ADM technology in public administration. ADM in public administration is a phenomenon that comes in a wide range of formats: from the use of automatic information processing for use as one part of basic administrative over semi-automated decision-making, to fully automated decision-making that uses AI to link information about facts to legal rules via machine learning.[20] While in theory a full spectrum of approaches is possible, and fully automated models have attracted a lot of attention,[21] in practice most forms of ADM are a type of hybrid system. As a prototype of what a hybrid process that would protect against many of the fears associated with ADM might look like, we introduce a novel solution, that we, for lack of a better term, call the 'administrative Turing test' (Section 11.4). This test could be used to *continually validate and strengthen* AI-supported decision-making. As the name indicates, it relies on comparing solely human and algorithmic decisions, and only allows the latter when a human cannot immediately tell the difference between the two. The administrative Turing test is an instrument to ensure that the existing (human) explanation requirement is met in practice. Using this test in ADM systems aims at ensuring the continuous quality of explanations in ADM and advancing what some research suggests is the

[20] There is indeed also a wide range of ways that an automated decision can take place. For an explanation of this, see the working version of this paper at section 3, http://ssrn.com/abstract=3402974.

[21] Perhaps most famous is O'Neil (n 3), but the debate on Technological Singularity has attracted a lot of attention; see, for an overview, Murray Shanahan, *The Technological Singularity* (MIT Press 2015).

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

best way to use AI for legal purposes – namely, in collaboration with human intelligence.[22]

11.2 EXPLANATION: CAUSAL VERSUS LEGAL

As mentioned previously, we focus on legal explanation – that is, a duty to give reasons/justifications for a legal decision. This differs from causal explainability, which speaks to an ability to explain the inner workings of that system beyond legal justification. Much of the literature on black-box AI has focused on the perceived need to open up the black box.[23] We can understand that this may be because it is taken for granted that a human is by default explainable, where algorithms in their many forms are not, at least in the same way. We propose that, perhaps counter-intuitively, that even if we take the blackest of boxes, it is the legal requirement of explanation in the form of sufficient reasons that matter for the protection of citizens. It is, in our view, the ability to challenge, appeal, and assess decisions against their legal basis, which ensures citizens of protection. It is not a feature of being able to look into the minutiae of the inner workings of a human mind (its neuronal mechanisms) or a machine (its mathematical formulas). The general call for explainability in AI – often conflated with complete transparency – is not required for the contestation of the decision by a citizen. This does not mean that we think that the quest for transparent ADM should be abandoned. On the contrary, we consider transparency to be desirable, but we see this as a broader and more general issue that links more to overall trust in AI technology as a whole[24] rather than something that is necessary to meet the explanation requirement in administrative law. The requirement of explanation for administrative decisions can be found, in one guise or another, in most legal systems. In Europe, it is often referred to as the 'duty to give reasons' – that is, a positive obligation on administrative agencies to provide an explanation ('begrundelse' in Danish, 'Begründung' in German, and 'motivation' in French) for their decisions. The explanation is closely linked to the right to legal remedies. Some research indicates that its emergence throughout history has been driven by the need to enable the citizen affected by an administrative decision to effectively challenge it before a court of

[22] See Saul Levmore and Frank Fagan, 'The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion' (2019) 93 *Southern California Law Review*.

[23] See, for example, Riccardo Guidotti et al., 'A Survey of Methods for Explaining Black Box Models' (2018) 51 *ACM Computing Surveys (CSUR)* 1. Similarly, Cobbe (n 2), who makes a distinction between 'how' and 'why' a decision was made, says 'just as it is often not straightforward to explain *how* an ADM system reached a particular conclusion, so it is also not straightforward to determine *why* that system reached that conclusion'. Our point is that these are the wrong questions to ask, because even in a human non-ADM system, we will never know 'why that system reached that conclusion'. We cannot know. What we can *do*, however, is to judge whether or not the explanation given was sufficiently accurate and sufficient under the given legal duty to give reasons.

[24] Amina Adadi and Mohammed Berrada, 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)' (2018) 6 *IEEE Access* 52138.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

law.[25] This, in turn, required the provision of sufficient reasons for the decision in question: both towards the citizen, who as the immediate recipient should be given a chance to understand the main reasoning behind the decision, and the judges, who will be charged with examining the legality of the decision in the event of a legal challenge. The duty to give reasons has today become a self-standing legal requirement, serving a multitude of other functions beyond ensuring effective legal remedies, such as ensuring better clarification, consistency, and documentation of the decisions, self-control of the decision-makers, internal and external control of the administration as a whole, as well as general democratic acceptance and transparency.[26]

The requirement to provide an explanation should be understood in terms of the law that regulates the administrative body's decision in the case before it. It is not a requirement that *any* kind of explanation must or should be given but rather a *specific kind* of explanation. This observation has a bearing on the kind of explanation that may be required for administrative decision-making relying on algorithmic information analysis as part of the process towards reaching a decision. Take, for instance, our example of Parent A. An administrative body issues a decision to Parent A in the form of a rejection explaining that the illness the child suffers from does not qualify as *serious* within the meaning of the statute. The constituents of this explanation would generally cover a reference to the child's disease and the qualifying components of the category of *serious illness* being applied. This could be, for example, a checklist system of symptoms or a reference to an authoritative list of formal diagnoses that qualify combined with an explanation of the differences between the applicant disease and those categorised as applicable under the statute. In general, the decision to reject the application for compensation of lost income would explain the legislative grounds on which the decision rests, the salient facts of the case, and the most important connection points between them (i.e., the discretionary or interpretive elements that are attributed weight in the decision-making

[25] Uwe Kischel, *Die Begründung: Zur Erläuterung Staatlicher Entscheidungen Gegenüber Dem Bürger*, vol 94 (Mohr Siebeck 2003) 32–34.

[26] Franz-Joseph Peine and Thorsten Siegel, *Allgemeines Verwaltungsrecht* (12th ed., C.F. Müller2018) 160, mn. 513; Schweickhardt, Vondung, and Zimmermann-Kreher (eds), *Allgemeines Verwaltungsrecht* (10th ed., Kohlhammer 2018) 586–588; Kischel (n 25) 40–65; H. C. H. Hofmann, G. C. Rowe, and A. H. Türk, *Administrative Law and Policy of the European Union* (Oxford University Press 2011), 200–202; CJEU, *Council of the European Union* v. *Nadiany Bamba*, 15 November 2012, Case C-417 / 11, para. 49; N. Songolo, 'La motivation des actes administratifs', 2011, www.village-justice.com/articles/motivation-actes-administratifs,10849.html; J.-L. Autin, La motivation des actes administratifs unilatéraux, entre tradition nationale et évolution des droits européens 'RFDA' 2011, no. 137–138, 85–99. We do not engage in a deeper analysis of the underlying rationale for the existence of the requirement to provide an explanation, as this is not the aim of our chapter. For this discussion in administrative law, see Joana Mendes, 'The Foundations of the Duty to Give Reasons and a Normative Reconstruction' in Elizabeth Fisher, Jeff King, and Alison Young (eds), *The Foundations and Future of Public Law* (Oxford University Press 2020).

process).[27] It is against this background that the threshold for what an explanation requires should be understood.

In a human system, at no point would the administrative body be required to describe the neurological activity of the caseworkers that have been involved in making the decision in the case. Nor would they be required to provide a psychological profile and biography of the administrator involved in making the decision, giving a history of the vetting and training of the individuals involved, their educational backgrounds, or other such information, to account for all the inputs that may have been explicitly or implicitly used to consider the application. When the same process involves an ADM system, must the explanation open up the opaqueness of its mathematical weighting? Must it provide a technical profile of all the inputs into the system? We think not. In the case of a hybrid system with a human in the loop, must the administrators set out – in detail – the electronic circuits that connect the computer keyboard to the computer hard drive and the computer code behind the text-processing program used? Must it describe the interaction between the neurological activity of the caseworker's brain and the manipulation of keyboard tabs leading to the text being printed out, first on a screen, then on paper, and finally sent to the citizen as an explanation of how the decision was made? Again, we think not.

The provided examples illustrate the point that causal explanation can be both insufficient and superfluous. Even though it may be empirically fully accurate, it does not necessarily meet the requirement of *legal* explanation. It gives an explanation – but it does likely not give the citizen the explanation he or she is looking for. The problem, more precisely, is that the explanation provided by causality does not, in itself, normatively connect the decision to its legal basis. It is, in other words, not possible to see the *legal reasoning* leading from the facts of the case and the law to the legal decision, unless, of course, such legal reasoning is explicitly coded in the algorithm. The reasons that make information about the neurological processes inside the brains of decision-makers irrelevant to the legal explanation requirement are the same that can make information about the algorithmic processes in an administrative support system similarly irrelevant. This is not as controversial of a position as it might seem on first glance.

Retaining the existing human standard for explanation, rather than introducing a new standard devised specifically for AI-supported decision-making, has the extra advantage that the issuing administrative agency remains fully responsible for the decision no matter how it has been produced. From this also follows that the administrative agency issuing the decision can be queried about the decision in ordinary language. This then assures a focus on the *rationale* behind the explanation being respected, even if the decision has been arrived at through some algorithmic

[27] Making sure that the connection relies on 'clean' data is obviously very important, but it is a separate issue that we do not touch on in this chapter. For a discussion of this issue in regards to AI-supported law enforcement, see Richardson, Schultz, and Crawford (n 11).

calculation that is not transparent. If the analogy is apt in comparing algorithmic processes to human neurology or psychological history, then requiring algorithmic transparency in legal decisions that rely on AI-supported decision-making would fail to address the explanation requirement at the right level. Much in line with Rahwan et al., who argue for a new field of research – the study of machine behaviour akin to human behavioural research[28] – we argue that the inner workings of an algorithm are not what is in need of explanation but, rather, the human interaction with the *output* of the algorithm and the biases that lie in the *inputs*. What is needed is not that algorithms should be made more transparent, but that the standard for intelligibility should remain undiminished.

11.3 EXPLANATION: THE LEGAL STANDARD

A legal standard for the explanation of administrative decision-making exists across all main jurisdictions in Europe. We found, when looking at different national jurisdictions (Germany, France, Denmark, and the UK) and regional frameworks (EU law and European Human Rights law), that explanation requirements differ slightly among them but still hold as a general principle that never requires the kind of full transparency advocated for. While limited in scope, the law we investigated includes a variety of different legal cultures across Europe at different stages of developing digitalised administrations (i.e., both front-runners and late-comers in that process). They also diverge on how they address explanation: in the form of a general duty in administrative law (Denmark and Germany) or a patchwork of specific legislation and procedural safeguards, partly developed in legal practice (France and the UK). Common for all jurisdictions is that the legal requirement put on administrative agencies to provide reasons for their decisions has a threshold level (minimum requirement) that is robust enough to ensure that if black box technology is used as part of the decision-making process, recipients will not be any worse off than if decisions were made by humans only. In the following discussion, we will give a brief overview of how the explanation requirement is set out in various jurisdictions.[29]

In Denmark, The Danish Act on Public Administration contains a section on explanation (§§22-24).[30] In general, the explanation can be said to entail that the citizen to whom the decision is directed must be given sufficient information about the grounds of the decision. This means that the explanation must fully cover the decision and not just explain parts of the decision. The explanation must also be truthful and in that sense correctly set forth the grounds that support the decision. Explanations may be limited to stating that some factual requirement in the case is

[28] See Iyad Rahwan et al., 'Machine Behaviour' (2019) 568 *Nature* 477.

[29] For a longer detailed analysis, see the working paper version of this chapter: http://ssrn.com/abstract=3402974.

[30] The full text at www.retsinformation.dk/forms/r0710.aspx?id=161411#Kap6.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

not fulfilled. For example, in our parent A example, perhaps a certain age has not been reached, a doctor's certificate is not provided, or a spouse's acceptance has not been delivered in the correct form. Explanations may also take the form of standard formulations that are used frequently in the same kind of cases, but the law always requires a certain level of concreteness in the explanation that is linked to the specific circumstances of the case and the decision being made. It does not seem to be possible to formulate any specific standards in regards to how deep or broad an explanation should be in order to fulfil the minimum requirement under the law. The requirement is generally interpreted as meaning explanations should reflect the most important elements of the case relevant to the decision. Similarly, in Germany, the general requirement to explain administrative decisions can be found in the Administrative Procedural Code of 1976.[31] Generally speaking, every written (or electronic) decision requires an explanation or a 'statement of grounds'; it should outline the essential factual and legal reasons that gave rise to the decision.

Where there was not a specific requirement for explanation,[32] we found – while perhaps missing the overarching general administrative duty – a duty to give reasons as a procedural safeguard. For example, French constitutional law does not by itself impose a general duty on administrative bodies to explain their decisions. Beyond sanctions of a punitive character, administrative decisions need to be reasoned, as provided by a 1979 statute[33] and the 2016 Code des Relations entre le Public et l'Administration (CRPA). The CRPA requires a written explanation that includes an account of the legal and factual considerations underlying the decision.[34] The rationale behind the explainability requirement is to strengthen transparency and trust in the administration, and to allow for its review and challenge before a court of law.[35] Similarly, in the UK, a recent study found, unlike many statements to the contrary and even without a *general* duty, in most cases, 'the administrative decision-maker being challenged [regarding a decision] was under a *specific statutory duty* to compile and disclose a specific statement of reasons for its decision'.[36] This research

[31] §39 VwVfG. Specialised regimes, e.g., for taxes and social welfare, contain similar provisions.

[32] We found that in neither France nor the UK is there a general duty for administrative authorities to give reasons for their decisions. For French law, see the decision by Conseil Constitutionnel 1 juillet 2004, no. 2004–497 DC ('les règles et principes de valeur constitutionnelle n'imposent pas par eux-mêmes aux autorités administratives de motiver leurs décisions dès lors qu'elles ne prononcent pas une sanction ayant le caractère d'une punition'). For UK law, see the decision by House of Lords in *R* v. *Secretary of State for the Home Department, ex parte Doody*, 1993 WLR 154 ('the law does not at present recognise a general duty to give reasons for an administrative decision').

[33] Loi du 11 juillet 1979 relative à la motivation des actes administratifs et à l'amélioration des relations entre l'administration et le public.

[34] Art. L211-5 ('La motivation exigée par le présent chapitre doit être écrite et comporter l'énoncé des considérations de droit et de fait qui constituent le fondement de la decision').

[35] N. Songolo, 'La motivation des actes administratifs, 2011', www.village-justice.com/articles/motivation-actes-administratifs,10849.html.

[36] Joanna Bell, 'Reason-Giving in Administrative Law: Where Are We and Why Have the Courts Not Embraced the "General Common Law Duty to Give Reasons"?' *The Modern Law Review* 9 http://

is echoed by Jennifer Cobbe, who found that 'the more serious the decision and its effects, the greater the need to give reasons for it'.[37]

In both the UK as well as the above countries, there are ample legislative safeguards that provide specific calls for reason giving. What is normally at stake is the *adequacy* of reasons that are given. As Marion Oswald has pointed out, the case law in the UK has a significant history in spelling out what is required when giving reasons for a decision.[38] As she recounts from *Dover District Council*, 'the content of [the duty to give reasons] should not in principle turn on differences in the procedures by which it is arrived at'.[39] What is paramount in the UK conception is not a differentiation between man and machine but one that stands by enshrined and tested principles of being able to mount a meaningful appeal, 'administrative law principles governing the way that state actors take decisions via human decision-makers, combined with judicial review actions, evidential processes and the adversarial legal system, are designed to counter' any ambiguity in the true reasons behind a decision.[40]

The explanation requirement in national law is echoed and further hardened in the regional approaches, where for instance Art. 41 of the Charter of Fundamental Rights of the European Union (CFR) from 2000 provides for a *right to good administration*, where all unilateral acts that generate legal consequences – and qualify for judicial review under Art. 263 TFEU – require an explanation.[41] It must 'contain the considerations of fact and law which determined the decision'.[42] Perhaps the most glaring difference that would arise between automated and non-automated scenarios is the direct application of Art. 22 of the General Data Protection Regulation (GDPR), which applies specifically to 'Automated individual decision making, including profiling.' Art. 22 stipulates that a data subject 'shall have the 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',[43] unless it is proscribed by law with 'sufficient

onlinelibrary.wiley.com/doi/abs/10.1111/1468-2230.12457 accessed 19 September 2019 original emphasis.

37 Cobbe (n 2) 648.

38 Marion Oswald, 'Algorithm-Assisted Decision-Making in the Public Sector: Framing the Issues Using Administrative Law Rules Governing Discretionary Power' (2018) 376 *Phil. Trans. R. Soc. A* https://ssrn.com/abstract=3216435.

39 *Dover District Council (Appellant)* v. *CPRE Kent (Respondent) CPRE Kent (Respondent)* v. *China Gateway International Limited (Appellant)* [2017] UKSC 79, para. 41. See, in particular, *Stefan* v. *General Medical Council* [1999] 1 WLR 1293 at page 1300G.

40 Oswald (n 38) 6.

41 Case C-370/07 *Commission of the European Communities* v. *Council of the European Union*, 2009, ECR I-08917, recital 42 ('which is justified in particular by the need for the Court to be able to exercise judicial review, must apply to all acts which may be the subject of an action for annulment').

42 Jürgen Schwarze, *European Administrative Law* (Sweet & Maxwell 2006) 1406.

43 Reg (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 Dir 95/46/EC (General Data Protection Regulation) 2016, Art. 22(1).

safeguards' in place,[44] or by 'direct consent.'[45] These sufficient safeguards range from transparency in the input phase (informing and getting consent) to the output-explanation phase (review of the decision itself).[46] The GDPR envisages this output phase in the form of external auditing through Data Protection Authorities (DPAs), which have significant downsides in terms of effectiveness and efficiency.[47] Compared to this, we find the explanation standard in administrative law to be much more robust, for it holds administrative agencies to a standard for intelligibility irrespective of whether they use ADM or not. Furthermore, under administrative law, the principle of 'the greater interference on the recipients life a decision has, the greater the need to give reasons in justification of the decision' applies. Furthermore, the greater the discretionary power of the decision maker, the more thorough the explanation has to be.[48] Focusing on the process by which a decision is made rather than the gravity of its consequences seems misplaced. By holding on to these principles, the incentive should be to develop ADM technology that can be used under this standard, rather than inventing new standards that fit existing technologies.[49]

ADM in public administration does not and should not alter existing explanation requirements. The explanation is not different now that it is algorithmic. The duty of explanation, although constructed differently in different jurisdictions, provides a robust foundation across Europe for ensuring that decision-making in public administration remains comprehensible and challengeable, even when ADM is applied. What remains is asking how ADM could be integrated into the decision-making procedure in the organisation of a public authority to ensure this standard.

11.4 ENSURING EXPLANATION THROUGH HYBRID SYSTEMS

Introducing a machine-learning algorithm in public administration and using it to produce *drafts* of decisions rather than final decisions to be issued immediately to citizens, we suggest, would be a useful first step. In this final section of the chapter, we propose an idea that could be developed into a proof of concept for how ADM could be implemented in public authorities to support decision-making.

In contemporary public administration, much drafting takes place using templates. ADM could be coupled to such templates in various ways. Different

[44] Ibid., Art. 22(2)b.

[45] Ibid., Art. 22(2)c.

[46] For a longer detailed analysis, see the working paper version of this chapter: http://ssrn.com/abstract=3402974.

[47] See Antoni Roig, 'Safeguards for the Right Not to Be Subject to a Decision Based Solely on Automated Processing (Article 22 GDPR)' (2017) 8(3) *European Journal of Law and Technology*.

[48] Schwarze (n 42) 1410.

[49] See also Zalnieriute, Moses, and Williams (n 2), who conclude (at p. 454) after conducting four case studies that only one system (the Swedish student welfare management system) succeeds in reaping benefits from automation while remaining sensitive to rule of law values. They characterize this as 'a carefully designed system integrating automation with human responsibility'.

templates require different kinds of information. Such information could be collected and inserted into the template automatically, as choices are made by a human about what kind of information should be filled into the template. Another way is to rely on automatic legal information retrieval. Human administrators often look to previous decisions of the same kind as inspiration for deciding new cases. Such processes can be labour intensive, and the same public authority may not all have the same skills in finding a relevant, former decision. Natural Language Processing technology may be applied to automatically retrieve relevant former decisions, if the authority's decisions are available in electronic form in a database. This requires, of course, that the data the algorithm is learning from is sufficiently large and that the decisions in the database are generally considered to still be relevant 'precedent'[50] for new decisions. Algorithmically learning from historical cases and reproducing their language in new cases by connecting legal outcomes to given fact descriptions is not far from what human civil servants would do anyway: whenever a caseworker is attending to a new case, he or she will seek out former cases of the same kind to use as a compass to indicate how the new case should be decided.

One important difference between a human and an algorithm is that humans have the ability to respond more organically to past cases because they have a broader horizon of understanding: They are capable of contextualizing the understanding of their task to a much richer extent than algorithms, and humans can therefore adjust their decisions to a broader spectrum of factors – including ones that are hidden from the explicit legislation and case law that applies to the case at hand.[51] Resource allocation, policy signals, and social and economic change are examples of this. This human contextualisation of legal text precisely explains why new practices sometimes develop under the same law.[52]. Algorithms, on the other hand operate, without such context and can only relate to explicit texts. Hence they cannot evolve in the same way. Paradoxically, then, having humans in the legal loop serves the purpose of relativizing strict rule-following by allowing sensitivity to context.

This limited contextualization of algorithmic 'reasoning' will create a problem if *all* new decisions are drafted on the basis of a machine learning algorithm that reproduces the past, and if those drafts are only subjected to minor or no changes by

[50] We are well aware that such decisions do not formally have the character of precedent, what we refer to here is the de facto tendency in the administrative process to make new decisions that closely emulate earlier decisions of the same kind.

[51] Even deciding what former decisions are relevant to a new case can sometimes be a complex problem that requires a broader contextual understanding of law and society that is not attainable by algorithms.

[52] See also Carol Harlow and Richard Rawlings, 'Proceduralism and Automation: Challenges to the Values of Administrative Law' in E. Fisher, J. King, and A. Young (eds), *The Foundations and Future of Public Law (in Honour of Paul Craig)* (Oxford University Press 2019) (at 6 in the SSRN version) https://papers.ssrn.com/abstract=3334783, who note that 'Administrative Law cannot be static, and the list of values is not immutable; it varies in different legal orders and over time'.

its human collaborator[53]. Once the initial learning stage is finalized and the algorithm is used in output mode to produce decision drafts, then new decisions will be produced in part by the algorithm. One of two different situations may now occur: One, the new decisions are fed back into the machine-learning stage. In this case, a feedback loop is created in which the algorithm is fed its own decisions.[54] Or, two, the machine-learning stage is blocked after the initial training phase. In this case, every new decision is based on what the algorithm picked up from the original training set, and the output from the algorithm will remain statically linked to this increasingly old data set. None of these options are in our opinion optimal for maintaining an up-to-date algorithmic support system.

There are good reasons to think that a machine learning algorithm will only keep performing well in changing contexts (which in this case is measured by the algorithm's ability to issue usable drafts of a good legal quality) – if it is constantly maintained by fresh input which reflects those changing contexts. This can be done in a number of different ways, depending on how the algorithmic support system is implemented in the overall organization of the administrative body and its procedures for issuing decisions. As mentioned previously, our focus is on models that engage AI and human collaboration. We propose two such models for organizing algorithmic support in an administrative system that aim at issuing decisions that we think are particularly helpful because they address the need for intelligible explanations of the outlined legal standard.

In our first proposed model, the caseload in an administrative field that is supported by ADM assistance is randomly split into two loads, such that one load is fed to an algorithm for drafting and another load is fed to a human team, also for drafting. Drafts from both algorithms and humans are subsequently sent to a senior civil servant (say a head of office), who finalizes and signs off on the decisions. All final decisions are pooled and used to regularly update the algorithm used.

By having an experienced civil servant interact with algorithmic drafting in this way, and feeding decisions, all checked by human intelligence, back into the machine-learning process, the algorithm will be kept fresh with new original decisions, a percentage of which will be written by humans from scratch. The effect of splitting the caseload and leaving one part to through a 'human only' track is that the previously mentioned sensitivity to broader contextualization is fed back into the algorithm and hence allows a development in the case law that could otherwise not

[53] Research has identified a phenomenon known as *automation bias*. This is the propensity for humans to favour suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct. See Mary Cummings, 'Automation Bias in Intelligent Time Critical Decision Support Systems', *AIAA 1st Intelligent Systems Technical Conference* (2004); Asia J Biega, Krishna P Gummadi, and Gerhard Weikum, 'Equity of Attention: Amortizing Individual Fairness in Rankings', *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018). In implementing ADM in public administration, we follow this research by recommending processes that seek to reduce such bias.

[54] See O'Neil (n 3) for a discussion of the problem with feedback loops.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

happen. To use our Parent A example as an illustration: Over time, it might be that new diseases and new forms of handicaps are identified or recognized as falling under the legislative provision because it is being diagnosed differently. If every new decision is produced by an ADM system that is not updated with new learning on cases that reflect this kind of change, then the system cannot evolve to take the renewed diagnostic practices into account. To avoid this 'freezing of time', a hybrid system in which the ADM is constantly being surveyed and challenged is necessary. Furthermore, if drafting is kept anonymous, and all final decisions are signed off by a human, recipients of decisions (like our Parent A) may not know how his/her decision was produced. Still, the explanation requirement assures that recipients can at any time challenge the decision, by inquiring further into the legal justification.[55] We think this way of introducing algorithmic support for administrative decisions could advance many of the efficiency and consistency (equality) gains sought by introducing algorithmic support systems, while preserving the legal standard for explanation.

An alternative method – our second proposed model – is to build into the administrative system itself a kind of continuous administrative Turing test. Alan Turing, in a paper written in 1950,[56] sought to identify a test for artificial intelligence. The test he devised consisted of a setup in which (roughly explained) two computers were installed in separate rooms. One computer was operated by a person; the other was operated by an algorithmic system (a machine). In a third room, a human 'judge' was sitting with a third computer. The judge would type questions on his computer, and the questions would then be sent to both the human and the machine in the two other rooms for them to read. They would then in turn write replies and send those back to the judge. If the judge could not identify which answers came from the person and which came from the machine, then the machine would be said to have shown the ability to think. A model of Turing's proposed experimental setup is seen in Figure 11.1:

Akin to this, an administrative body could implement algorithmic decision support in a way that would imitate the setup described by Turing. This could be done by giving it to both a human administrator and an ADM. Both the human and the ADM would produce a decision draft for the same case. Both drafts would be sent to a human judge (i.e., a senior civil servant who finalizes and signs off on the decision). In this setup, the human judge would not know which draft came from the ADM and which came from the human,[57] but would proceed to finalize the decision based on which draft was most convincing for deciding the case and

[55] Whether recipients can or should be able to demand insight into the underlying neurological or algorithmic computations of caseworkers (human or robotic) is a separate question that we do not seek to answer here. Suffice it to say there may be many reasons why a human might ask for an explanation, including not caring what the justification is but simply wanting a change of outcome.

[56] A. M. Turing, 'Computing Machinery and Intelligence' (1950) 49 *Mind* 433–460.

[57] Formats for issuing drafts could also be formalized so as to reduce the possibility of guessing merely by recognizing the style of the drafter's language.

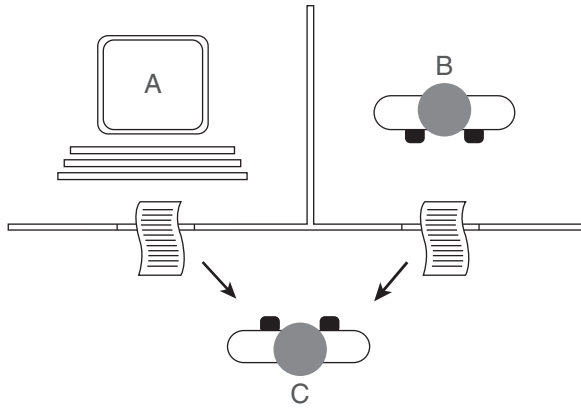

FIGURE 11.1 Turing's experimental setup (Source: https://en.wikipedia.org/wiki/Turing_test)

providing a satisfactory explanation to the citizen. This final decision would then be fed back to the data set from which the ADM system learns.

The two methods described previously are both hybrid models and can be used either alone or in combination to assure that ADM models are implemented in a way that is both productive, because drafting is usually a very time-consuming process and safe (even if not mathematically transparent) because there is a human overseeing the final product and a continuous human feedback to the data set from which the ADM system learns. Moreover, using this hybrid approach helps overcome the legal challenges that a fully automated system would face from both EU law (GDPR) and some domestic legislation.

11.5 CONCLUSION

Relying on the above models keeps the much-sought-after 'human in the loop' and does so in a way that is systematic and meaningful because our proposed models take a specific form: they are built around the idea of continuous human-AI collaboration in producing explainable decisions. Relying on this model makes it possible to develop ADM systems that can be introduced to enhance the effectiveness, consistency (equality) without diminishing the quality of explanation. The advantage of our model is that it allows ADM to be continuously developed and fitted to the legal environment in which it is supposed to serve. Furthermore, such an approach may have further advantages. Using ADM for legal information retrieval allows for analysis across large numbers of decisions that have been handed down across time. This could grow into a means for assuring better detection of hidden biases and other structural deficiencies that would otherwise not be discoverable. This approach may help allay the fears of the black box.

https://doi.org/10.1017/9781108914857.012 Published online by Cambridge University Press

In terms of control and responsibility, our proposed administrative Turing test allows for a greater scope of review of rubber stamp occurrences by being able to compare differences in pure human and pure machine decisions by a human arbiter. Therefore the model may also help in addressing the concern raised about 'retrospective justifications'.[58] Because decisions in the setup we propose are produced in collaboration between ADM and humans, the decisions issued are likely to be more authentic than either pure ADM or pure human decision-making, since the use of ADM allows for a more efficient and comprehensive inclusion of existing decision-making practice as inputting the new decision-making through automated information retrieval and recommendation. With reference to human dignity, our proposed model retains human intelligibility as the standard for decision-making. The proposed administrative Turing model also continually adds new information into the system, and undergoes a level of supervision that can protect against failures that are frequently associated with ADM systems. Applying the test developed in this chapter to develop a proof of concept for the implementation of ADM in public administration today is the most efficient way of overcoming the weaknesses of purely human decision-making tomorrow.

ADM does not solve the inequalities built into our societal and political institutions, nor is it their original cause. There are real questions to be asked of our systems, and we would rather not bury those questions with false enemies. To rectify those inequalities, we must be critical of our human failings and not hold hostage the principles we have developed to counter injustice. If those laws are deficient, it is not the fault of a new technology. We are, however, aware that this technology can not only reproduce but even heighten injustice if it is used thoughtlessly. But we would also like to flag that the technology offers an opportunity to bring legal commitments like the duty of explanation up to a standard that is demanded by every occurrence of injustice: a human-based standard.

[58] Cobbe remarks that black box technology that 'their inexplicability is therefore a serious issue' and therefore decisions issued by such systems will likely not pass judicial review. She then adds that 'some public bodies may attempt to circumvent this barrier by providing retrospective justifications'. She flags that Courts and reviewers should be 'aware of this risk and should be prepared to exercise the appropriate level of scrutiny . . . against such justifications.' Cobbe (n 2) 648.