Towards Intelligent Regulation of Artificial Intelligence

Miriam C BUITEN*

Artificial intelligence (AI) is becoming a part of our daily lives at a fast pace, offering myriad benefits for society. At the same time, there is concern about the unpredictability and uncontrollability of AI. In response, legislators and scholars call for more transparency and explainability of AI. This article considers what it would mean to require transparency of AI. It advocates looking beyond the opaque concept of AI, focusing on the concrete risks and biases of its underlying technology: machine-learning algorithms. The article discusses the biases that algorithms may produce through the input data, the testing of the algorithm and the decision model. Any transparency requirement for algorithms should result in explanations of these biases that are both understandable for the prospective recipients, and technically feasible for producers. Before asking how much transparency the law should require from algorithms, we should therefore consider if the explanation that programmers could offer is useful in specific legal contexts.

1. Introduction

"Any sufficiently advanced technology is indistinguishable from magic"
Arthur C Clarke[1]

Artificial intelligence (AI) is becoming more enmeshed in our daily lives at a fast pace. AI applications guide a vast array of decisions that up until recently only extensively trained humans could perform. This has allowed for remarkable improvements in numerous fields. For instance, AI applications are analysing images to detect potentially cancerous cells,[2] helping to predict where and when the next big earthquake will strike[3]

* The author gratefully acknowledges financial support from Deutsche Forschungsgemeinschaft (DFG) through CRC TR 224 and wishes to thank an anonymous referee and the participants of the McGill Faculty of Law Cipp/Lallemand Seminar. Author email: buiten@uni-mannheim.de.

[1] Arthur C Clarke, *Profiles of the Future: An Inquiry into the Limits of the Possible* (Harper & Row, Revised edition 1973).

[2] ARM Al-shamasneh and UHB Obaidellah, "Artificial intelligence techniques for cancer detection and classification: review study" (2017) 13 Eur Sci J 342.

[3] T Fuller and C Metz, "A.I. Is Helping Scientists Predict When and Where the Next Big Earthquake Will Be" (*New York Times*, 26 October 2018) < www.nytimes.com/2018/10/26/technology/earthquake-predictions-artificial-intelligence.html > accessed 14 February 2019.

European Journal of Risk Regulation, 10 (2019), pp. 41–59 doi:10.1017/err.2019.8

and allowing for the development of companion robots supporting overburdened caretaking staff.[4]

We have also observed several cases in which AI systems produced poor outcomes. We witnessed the first fatal accident involving an autonomous vehicle, when a Tesla car in self-driving mode killed a pedestrian in 2016.[5] In 2017, Microsoft's chatting bot *Tay* had to be shut down after 16 hours because it became racist, sexist, and denied the Holocaust.[6] These and other examples have fuelled a variety of concerns about the accountability, fairness, bias, autonomy, and due process of AI systems.[7]

In response to the increasing relevance of AI in society, pressure is mounting to make AI applications more transparent and explainable. The 2016 White House report on AI suggests that many of the ethical issues related to AI can be addressed through increasing transparency.[8] The report calls on researchers to "develop systems that are transparent, and intrinsically capable of explaining the reasons for their results to users".[9] The UK House of Lords Artificial Intelligence Committee has recently suggested that in order to achieve trust in AI tools, deploying any AI system that could have a substantial impact on an individual's life is only acceptable if it can generate a full and satisfactory explanation for the decisions it will take.[10] The European Parliament in its 2016 report on AI notes that "it should always be possible to supply the rationale behind any decision taken with the aid of AI that can have a substantive impact on one or more persons' lives" and "to reduce the AI system's computations to a form comprehensible by humans".[11] Moreover, the General Data Protection Regulation (GDPR) lays down a right for a data subject to receive "meaningful information about the logic involved" if not only information is collected about them, but also profiling takes place.[12]

---

[4] K Purvis, "Meet Pepper the robot – Southend's newest social care recruit" (*The Guardian*, 16 October 2017) < www.theguardian.com/social-care-network/2017/oct/16/pepper-robot-southend-social-care-recruit > accessed 14 February 2019.

[5] K Deamer, "What the First Driverless Car Fatality Means for Self-Driving Tech" (*Scientific American*, 1 July 2016) < www.scientificamerican.com/article/what-the-first-driverless-car-fatality-means-for-self-driving-tech/> accessed 14 February 2019; Tesla Motors statement (30 June 2016) < www.teslamotors.com/en_GB/blog/tragic-loss > accessed 14 February 2019.

[6] S Perez, "Microsoft Silences its New A.I. Bot Tay, After Twitter Users Teach It Racism" (*Techcrunch*, 24 March 2016) < techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism > accessed 14 February 2019.

[7] F Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information* (Harvard University Press 2015); M Ziewitz, "Governing algorithms: Myth, mess, and methods" (2016) 41(1) ST&HV 3, 4; N Bostrom, "Ethical issues in advanced artificial intelligence" in S Schneider (ed), *Science Fiction and Philosophy: From Time Travel to Superintelligence* (John Wiley & Sons 2010); D Amodei et al, "Concrete problems in AI safety" (2016) < arxiv.org/abs/1606.06565 > accessed 14 February 2019; D Sculley et al, "Machine learning: The high-interest credit card of technical debt" (2014) < research.google.com/pubs/archive/43146.pdf > accessed 14 February 2019.

[8] Executive Office of the President, *Artificial intelligence, automation and the economy* (2016). See also C Cath et al, "Artificial Intelligence and the 'Good Society': the US, EU, and UK approach" (2018) Sc Eng Ethics 505, 511.

[9] National Science and Technology Council Networking and Information Technology. Networking and Information Technology Research and Development Subcommittee, *The national artificial intelligence research and development strategic plan* (2016) 28.

[10] UK House of Lords Artificial Intelligence Committee, *AI in the UK: ready, willing and able?* (HL Paper 100 2018) para 105.

[11] European Parliament Committee on Legal Affairs, *Civil law rules on robotics* (2015/2103 (INL)) 10.

[12] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, Arts 12, 14 and 22.

In this article, I ask what transparency of AI actually means, as a starting point for evaluating whether requiring more transparency by law is feasible and useful. Any AI regulation, such as a transparency requirement, would need to define AI. There is, however, no single definition of AI. Various definitions characterise AI by the autonomy of systems or its "human-like" intelligent outcomes, without clearly demarcating the boundaries of AI. Essentially, AI is in the eye of the beholder. This is understandable for a rapidly evolving field of technology, but makes AI an unsuitable basis for regulation. Moreover, the opacity of the concept reinforces concerns about the uncontrollability of new technologies: we fear what we do not know. Whereas concerns of unknown risks of new technologies are valid, this does not mean that AI is completely unknowable. AI, in other words, is not magic.

I propose an approach that looks beyond AI as an undecipherable "black box". AI applications generally rely on algorithms. The exact workings of these algorithms may be inscrutable for the public. Nevertheless, we can study the main flaws and biases that algorithms present in order to get more grip on the risks that advanced automated systems present. Once we understand where algorithms may go wrong, we can consider what it means to require transparency of algorithms and whether this is practicable and desirable. In this article, I analyse how mistakes in the input data or training data, testing phase and decision model may result in flawed or undesirable outcomes of algorithms.

A transparency requirement of algorithms would need to explain these possible flaws and risks in a way that is both understandable for the prospective recipients, and technically feasible for producers. I consider legal cases concerning fundamental rights and civil liability, finding that effectively informing courts would require considerable engineering effort of algorithm producers, and potentially reduce algorithms' accuracy. We need to balance these costs against the utility of explanations for particular contexts. Moreover, regarding any transparency requirement we need to ask if a technically feasible explanation is going to be useful in practice.

II. AI AS A REGULATORY TARGET

1. The difficulty of defining AI: in search of a legal definition

A legal transparency requirement for AI – and in fact any regulatory regime for AI – would have to define AI. However, there does not yet appear to be any widely accepted definition of AI even among experts in the field.[13] Definitions of intelligence vary widely and may focus on AI as a field of research, the autonomy of advanced systems, a comparison with human intelligence, or the technologies and applications of AI.

John McCarthy first coined the term artificial intelligence in 1956,[14] defining AI as "the science and engineering of making intelligent machines, especially intelligent computer programs".[15] For the purpose of regulation, such a circular definition of AI is of little functional use.

[13] MU Scherer, "Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, And Strategies" (2016) 29(2) JOLT 354, 359. See also J McCarthy, "What is Artificial Intelligence?" (12 November 2007), < www-formal.stanford.edu/jmc/whatisai.pdf > accessed 14 February 2019.

[14] See further G Press, "Artificial Intelligence (AI) Defined" (*Forbes* 27 August 2017), < www.forbes.com/sites/gilpress/2017/08/27/artificial-intelligence-ai-defined/#3453304d7661 > accessed 14 February 2019.

[15] McCarthy, supra, note 13, 1.

Other definitions of AI focus on a certain degree of autonomy exhibited in advanced systems. Scherer views AI's ability to act autonomously as the most obvious feature of AI that separates it from earlier technologies.[16] As Gasser et al put it, AI applications often influence their environment or human behaviour and evolve dynamically in ways that are at times unforeseen by the systems' designers.[17] The problem with this definition for the purposes of law is that our views of what constitutes an autonomous or unforeseen decision by a computer are likely to change over time. What humans perceived as unpredictable behaviour (or even intelligent behaviour, as I will discuss below) of a computer 30, 20 or even 10 years ago may be considered "nothing special" today. Illustrative is a definition of AI by Cole in 1990, describing the class of programs that "emulate knowledgeable manipulation of the real world". Cole notes that "AI involves sensing, reasoning, and interaction within the real world, or at least a portion thereof; the effect of the program's outputs may become part of the program's later inputs".[18] Today, we see a myriad of algorithms whose outputs become part of later inputs: even relatively simple learning algorithms may do this. We might nevertheless require more than this learning ability to classify an application as AI. In short, definitions of AI focusing on the level of autonomy of the application may be interpreted differently depending on whom you ask, and when you ask it. Whereas the notion of autonomy may be a relevant factor for various legal problems surrounding AI,[19] it may be too subjective to serve as a basis for a legal definition.

Other definitions relate AI to human intelligence or to human characteristics, including consciousness, self-awareness, language use, the ability to learn, the ability to abstract, the ability to adapt, and the ability to reason.[20] The Merriam-Webster dictionary defines AI as "[a] branch of computer science dealing with the simulation of intelligent behavior in computers," or "[t]he capability of a machine to imitate intelligent human behaviour".[21] Russell and Norvig present eight different definitions of AI organised into four categories: thinking humanly, acting humanly, thinking rationally, and acting rationally.[22] As McCarthy puts it, there is no solid definition of intelligence that does not depend on relating it to human intelligence. The reason is that "we cannot yet characterize in general what kinds of computational procedures we want to call intelligent".[23] Essentially, we may thus characterise applications as AI when they produce results that we perceive as equivalent to human intelligence and therefore qualify as artificial intelligence. In other words, labelling AI has much to do with how intelligent we perceive the outcomes to be, presenting similar problems as defining AI in terms of "autonomy". Putting it bluntly, it means that AI is what we call AI. This definition is both circular and subjective, making it unsuitable as a basis for laws and regulations.

---

16 Scherer, supra, note 13, 363.

17 U Gasser and VAF Almeida, "*A layered model for AI governance*" (2017) 6 IEEE Internet Computing 58, 2.

18 GS Cole, "Tort Liability for Artificial Intelligence and Expert Systems" (1990) 10(2) Computer/Law Journal 127.

19 See further section I.2 below.

20 Scherer, supra, note 13, 360.

21 Merriam-Webster Online Dictionary, "Artificial Intelligence" < www.merriam-webster.com/dictionary/artificial%20intelligence > accessed 14 February 2019.

22 SJ Russell and P Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall 2010) 2.

23 McCarthy, supra, note 13.

Yet other definitions use AI as an umbrella term to refer to myriad advanced technologies and applications. The English Oxford Living Dictionary defines AI as "[t]he theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages".[24] Similarly, Stone et al consider AI to be a set of applications and sub-disciplines with various applications, such as speech recognition, health diagnostic systems, self-driving cars and computer vision to attention and memory.[25] These definitions illustrate that AI as a regulatory concept would probably cover a wide array of applications and uses, some of which may already be regulated in other ways. Introducing a regulatory regime to deal with the problems of all these AI applications would be impossibly wide in scope.[26]

Some legal commentators focus on robots rather than on AI. Calo defines robots as artificial, mechanical objects that take the world in, process what they sense, and in turn act upon the world to at least some degree.[27] Calo notes that this is just a technical definition, emphasising that for policy discourse it is not the precise architecture of robots that is important, but the possibilities and experience it generates and circumscribes. Others, such as Balkin, do not distinguish sharply between robots and AI agents, as the boundaries between the two technologies may increasingly blur and we do not yet know all the ways in which technology will be developed and deployed.[28] The term "robotics", in other words, is as open to interpretation as the concept of AI.

In sum, there does not appear to be a clear definition of AI. The various definitions of AI used in the literature may be helpful to understand AI, but are unsuitable as a basis for new laws. While some uncertainty may be inherent to new technologies, it is problematic to centre laws and policies around the opaque concept of AI.

2. Problems for laws and regulations

The concern of AI is that it presents new and unknown risks with which current laws may not be able to cope. We may thus need new rules to mitigate the risks of AI. One proposal is to introduce transparency requirements for AI.[29] Such a transparency requirement would be difficult to implement in practice, if we are not sure what its scope is. The lack of a clearly defined scope would leave the industry uncertain whether its activities are covered by the regulation, and would leave the definition of AI to the courts. The definitions of AI discussed above are, however, context-sensitive and time-varying in character.

Take the approach of defining AI in terms of a system's autonomy. This raises the question of what it means for an application to act autonomously. Scherer uses the example of autonomous cars and automated financial advisers that can perform complex

[24] Oxford Dictionaries, "Artificial Intelligence" <en.oxforddictionaries.com/definition/artificial_intelligence> accessed 22 October 2018.

[25] P Stone et al, "Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence" (2016) <ai100.stanford.edu/2016-report> accessed 20 September 2018.

[26] C Reed, "How should we regulate artificial intelligence?" (2018) Phil Trans R Soc A 1.

[27] R Calo, "Robotics and the Lessons of Cyberlaw" (2015) 103 Cal L Rev 513, 529.

[28] JM Balkin, "The Path of Robotics Law" (2015) 6 Cal L Rev 45, 51.

[29] See eg European Parliament, supra, note 11, and White House Report, supra, note 8.

tasks "without active human control or even supervision".[30] Let us consider the example of cars. Most would agree that a regular car is not autonomous, but some would say, like Scherer, that a self-driving car is. What about cars that have some autonomous-driving features, such as automatic distance keeping? Do they have sufficient autonomy to be considered AI, and should this distinction have any legal implications? If we were to require more transparency from AI, what would be the implications for producers of regular cars, partly autonomous and fully autonomous cars? Who would be subject to the requirement, and what would the requirement entail concretely? The problem, of course, would not be limited to cars, but to all systems that may be called AI. Considerable uncertainty would ensue on which systems and applications fall within the scope of such a rule.

Using intelligence to define AI presents similar problems. Capabilities we consider intelligent today may not be seen as such in the future. The AI literature from the 1980s is illustrative, referring to a system's ability to learn from experience as intelligent.[31] Systems capable of playing chess, for instance, were seen as AI systems.[32] Today, chess is considered as a relatively easy problem for a computer, given that the rules are known and the number of moves is finite. In 1996, IBM chess computer Deep Blue defeated the then world champion Gary Kasparov.[33] In essence, while AI is sometimes defined as the capability of machines to perform tasks that require intelligence, our conception of what constitutes intelligent may change as machines acquire more capabilities. After all, how can the task require intelligence if a machine can do it?[34] Even if we overcome the problem of varying definitions over time, intelligence may be viewed very differently depending on whom you ask. In 1982, John Searle essentially took the position that any technology that is understandable is not intelligent. Computers may seem to understand concepts but actually do not, as they attach no meaning or interpretation to any of it.[35] On the other end of the spectrum are the views discussed above of intelligent behaviour as mimicking intelligent human behaviour, or being perceived as intelligent by humans.[36] The range of applications that would be considered AI would vary greatly depending on which approach is taken.

In some instances, features of AI such as autonomy or intelligence may help guide us in incorporating AI in our laws.[37] When asking how laws should respond to AI, however, I propose to take a step back and demystify this concept. The narrative of AI as an inscrutable concept may reinforce the idea that AI is an uncontrollable force shaping our societies. In essence, part of the problem is that we cannot control what we do not

[30] Scherer, supra, note 13, 363.

[31] RC Schank, "The Current State of AI: One Man's Opinion" (1983) 4(1) AI Magazine 3, 4.

[32] J Haugeland, *Artificial Intelligence: The Very Idea* (MIT Press 1989).

[33] See further B Bloomfield and T Vurdabakis, "IBM's Chess Players: On AI and Its Supplements" (2008) 24 Inf Soc 69.

[34] See also Editorial, "What is AI? And What Does It Have to Do with Software Engineering?" (1985) SE-11 IEEE Transactions on Software Engineering 1253.

[35] JR Searle, "The Myth of the Computer" (1982) The New York Review of Books.

[36] See section II.1 above.

[37] See section II.3 below.

understand. I therefore propose to "open the black box" of AI and try to identify the risks and "unknowns".

This is not to say that AI does not present risks, or even that all of these risks are knowable now. In fact, the majority of the policy challenges regarding AI boil down to the need to mitigate risks that are unknown or difficult to control. It is clear that the question of how to address these risks deserves attention. A first step in doing so is to make these risks concrete, by considering the underlying computational systems. Notwithstanding AI being difficult to understand even for experts, it is a misconception that AI is completely unknowable. AI is a product of human invention, and humans continuously improve upon its underlying technology, algorithms. Algorithms are the basis of the wide array of AI applications, which may vary widely in their sophistication and purpose.[38]

Since the technology of algorithms is what AI applications have in common, I propose that we focus the policy debate on this technology, starting by better understanding the risks associated with algorithms. Understanding these risks may be possible without full comprehension of the technology, allowing us to grasp the risks and flaws that advanced automated systems present.

3. Technology as a regulatory target

Advocating a policy approach that "lifts the veil" of AI and considers the underlying technology of algorithms instead, does not necessarily mean that we should tailor new laws and regulations around this technology. New technologies are frequently the source of legal questions, occasionally with a new legal field ensuing.[39] Literature on previous technologies has illustrated the potential problems of focusing on the technology, rather than trying to solve the problems surrounding the technology within the existing legal subjects.

For instance, cyberlaw scholarship considered the desirability and nature of internet regulation. At a 1996 cyber law conference, Judge Frank Easterbrook compared cyberlaw to the "law of the horse". He meant the attempt to understand the law of horses by studying cases concerning sales of horses, people kicked by horses, or veterinary care of horses.[40] Easterbrook's principal objections to cyberlaw comprised that neither horses nor computers were a useful lens through which to view the law. Instead, we ought to study the problems related to horses or computers through traditional legal areas such as tort law, contract law and property law. Lawrence Lessig rebutted Judge Easterbrook's critique of cyberlaw, arguing that by identifying the distinctive feature of cyberspace one could find the key to regulating technology.[41]

We can draw several lessons from this debate for the policy issues surrounding algorithms. First, we should acknowledge that legal experts from various fields of law

[38] I Giuffridaet al, "A Legal Perspective on the Trials and Tribulations of AI: How Artificial Intelligence, the Internet of Things, Smart Contracts, and Other Technologies Will Affect the Law" (2018) 68(3) Case W Res L Rev 747, 753.

[39] E Katsh, *The Electronic Media and the Transformation of Law* (Oxford University Press 1989).

[40] FH Easterbrook, "*Cyberspace and the Law of the Horse*" (1996) University of Chicago Legal Forum 207.

[41] L Lessig, "The Law of the Horse: What Cyberlaw Might Teach" (1999) 113 Harvard Law Review 501–502, 509. See also DM Ibrahim and D Gordon Smith, "Entrepreneurs on Horseback: Reflections on the Organization of Law" (2008) 40 Arizona Law Review 71, 78.

might already have solutions for problems presented by a new technology. Whereas a new technology may seem to create new challenges, these challenges may simply be a version of a problem that laws already address effectively. In these cases, studying traditional legal areas may be more fruitful to solve new problems than bringing together diverse scholars to consider "the law of the algorithm". At the same time, diverse scholars may benefit from exchanging ideas, so that they can improve their mutual understanding of the technology. This second lesson is reflected by Lessig's critique, which for AI implies that we need to understand how algorithms affect decision-making.

In other words, we should make sure we avoid re-inventing the intellectual wheel for each new technology,[42] while acknowledging the key, distinguishing features of this technology. As Bennett Moses points out, for the regulation of technology to be a useful subject for scholarly examination, there must be something unique about questions around the regulation of the technology that do not apply when considering regulation more generally.[43]

Calo specifies this argument in the context of AI, proposing that we should identify the "essential qualities" of the new technology. Next, we should ask how the law should respond to the problems posed by those essential qualities. The three characteristics that Calo proposes are embodiment, emergence and social valence.[44] An entity has embodiment when it can interact with the world physically. The entity's degree of emergence reflects the unpredictability in its interaction with the environment. Social valence refers to whether people treat AI agents as human beings.[45] Balkin finds that rather than speaking in terms of "essential qualities" of a new technology, we should focus on what features of social life the technology makes newly salient: "What problems does a new technology place in the foreground that were previously underemphasized or deemed less important?"[46] We could then try to identify analogies for the new problems of AI in our current laws, ensuring that the law provides a solution for these problems.[47]

These approaches tackle both aspects of the above-mentioned technology debate: seeking solutions in existing legal fields, while aiming to identify how new technologies may affect these legal fields. Identifying the salient issues of new technologies for the law may guide lawmakers and courts in finding solutions for problems presented by these technologies and their applications.

Courts and lawmakers may nevertheless need concrete knowledge or information of algorithms when faced with these salient issues of AI. For instance, a lawmaker would have to know more about the roots of unpredictability in advanced systems to be able to draft rules and guidelines for these systems. A court would need some understanding of the roots of unpredictability of advanced systems when being asked to determine who is

[42] IN Cofone, "Servers and Waiters: What Matters in the Law of A.I." (2018) 21 Stan Tech L Rev 167, 197; L Bennett Moses, "How to Think about Law, Regulation and Technology: Problems with 'Technology' as a Regulatory Target" (2013) 5(1) LIT 1, 19.

[43] Bennett Moses supra, note 42, 11.

[44] Calo, supra, note 27, 515.

[45] ibid, 532–49.

[46] Balkin, supra, note 28, 46–47.

[47] Cofone, supra, note 42.

liable for harm caused by such an unpredictable system, in the design of which multiple parties may have been involved. Courts and regulators will need to "open the black box of AI" and ask what aspects of the technology present risks, where these risks stem from, and whether these risks can be controlled or contained. Depending on the extent and nature of the risk, this information may guide lawmakers on whether and how to respond to algorithms. For instance, it will inform the question of whether or not to impose strict liability for algorithms. If, instead, a negligence rule were imposed, a duty of care would need to be translated into technical requirements or safeguards.

In short, given that the concerns surrounding AI boil down to mitigating and controlling risks, it is helpful to consider these risks in more detail. Concretely, I propose to do so by asking what algorithms are, and in which ways their decision-making may be unpredictable or uncontrollable.

III. Algorithms: how they work and where they may go wrong

1. A brief introduction to algorithms[48]

In its most basic form, an algorithm is a set of instructions or rules given to a computer to follow and implement. A simple, rule-based algorithm is an unambiguous specification of how to solve a class of problems. These problems may include ordering possible choices (prioritisation), categorising items (classification), finding links between items (association) and removing irrelevant information (filtering), or a combination of these.

More sophisticated, machine learning (ML) algorithms are designed to learn, meaning to modify their programming to account for new data.[49] By applying ML algorithms a computer, with the help of training data, can learn rules and build a decision model. The computer does not merely execute explicit instructions but is programmed to find patterns in the data, turning them into the instructions that the programmers would have otherwise had to write themselves. ML is thus an attractive alternative to manually constructing these programs.

ML is used in web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications.[50] ML algorithms may help solve a wide array of problems, ranging from predicting how capable a credit applicant is of repaying a loan, identifying and labelling objects in images and videos, classifying protein patterns in human cells, converting written text into spoken forms, classifying malware, and so forth.[51]

Improvements in mathematical formulas and the availability of computing power allow computer scientists to employ more complex models. This has led to the development of deep learning, a branch within ML that uses a particular class of algorithms: artificial neural networks. Neural networks are frameworks for many different ML algorithms to work together and process complex data inputs. ML

[48] For a more extensive explanation of algorithms, the different types and their applications, see eg P Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World* (Basic Books 2015).

[49] Giuffrida, supra, note 38, 753.

[50] P Domingos, "A few useful things to know about machine learning" (2012) 55(10) Communications of the ACM 78, 78.

[51] Kaggle "Competitions" < www.kaggle.com/competitions > accessed 14 February 2019.

algorithms that use neural networks generally do not need to be programmed with specific rules that define predictions from the input. The learning algorithm in the neural network instead is given many examples with both input ("questions") and output ("answers") given, to learn what characteristics of the input are needed to construct the correct output. The more training data the neural network processes, the more accurately the neural network can begin to process new, unseen inputs and successfully return the right results.[52]

Deep learning has allowed for breakthroughs in for example speech and image recognition. For instance, a deep learning system of Google and Stanford learned to categorise objects in images without a human ever defining or labelling these objects.[53] At the same time, it becomes increasingly difficult to understand how these systems find the solution as they become more complex. In essence, in advance of their use, developers are not able to predict or explain their functioning. Currently, research is being done into providing retrospective explicability of neural net decisions[54].

Summarising, the capabilities of applications relying on algorithms depend on the sophistication of the algorithm. Somewhere on the continuum between simple algorithms and deep learning systems, we reach a level of complexity, sophistication and unpredictability of the outcomes that we may classify as AI. The concerns associated with these complex and intractable systems, however, are not tied to the system being qualified as AI. Understanding how algorithms reach decisions and what risks this decision-making process may present requires us to learn more about where in this decision-making process an algorithm may be flawed or biased.

2. Identifying the roots of biases

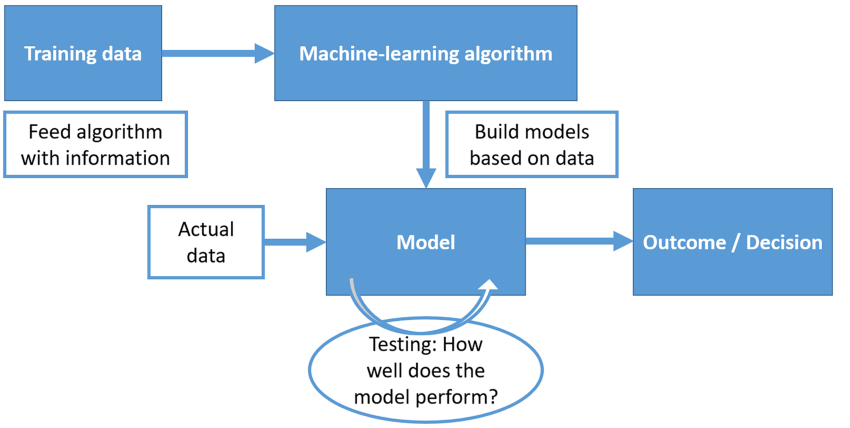
Many factors influence an algorithm's decision-making ability. Figure 1 provides a simplified graphical representation of the decision-making process of a machine-learning algorithm using the method of "supervised learning". First, training data is used to feed the algorithm with information. One could compare the training data with a set of questions with the answers already given. Next, the algorithm should derive a set of rules that allows construction of the output ("answers") from the input ("questions"). This set of rules makes up the decision model. The goal is to optimise the decision model as a predictor for the problem at hand: ideally, the decision model can construct or predict all output correctly from the input. In order to make sure that the decision model performs well in the environment where it is supposed to solve a problem, the model is tested with real-world input. Ultimately, the decision model can be used for actual problems: it is given data and produces outcomes or decisions.

We can now consider *data*, *testing* and the *decision model* as aspects of the algorithm's decision-making process, identifying errors that may lead to biased or harmful results.

[52] DeepAI, "Neural Network" (The Information, 14 March 2018) < deepai.org/machine-learning-glossary-and-terms/neural-network > accessed 14 February 2019.

[53] RD Hof, "Deep Learning" (MIT Technology Review) < www.technologyreview.com/s/513696/deep-learning/> accessed 14 February 2019.

[54] See further section IV.1.



a. Data

Machine learning uses computers to run predictive models that learn from existing data to forecast future behaviours, outcomes and trends.[55] Any ML algorithm, and its subsequent decisions, depends on the dataset that was used to train it. The more data it can access, the better its predictive model can be.

Large quantities of data alone are not enough, however: the quality of this data also affects the validity, accuracy and usefulness of the outcomes generated by the algorithm.[56] If key data is withheld by design or chance, the algorithm's performance might become very poor.[57] The often-used implicit assumption that once we collect enough data, algorithms will not be biased, is not justified.[58] Bias can arise in algorithms in several ways. First, the data we have collected may have been preferentially sampled, and therefore the data sample itself is biased.[59] Second, bias can arise because the collected data reflects existing societal bias.[60] To the extent that society contains inequality, exclusion or other traces of discrimination, so too will the data.[61] For instance, differences in arrest rates across racial groups may be replicated by an algorithm calculating recidivism risk.[62] Another example could be underrepresentation of women in particular jobs, from which a hiring algorithm may derive the rule that men

[55] J Sanito et al, "Deep Learning Explained" (EDX) < www.edx.org/course/deep-learning-explained-microsoft-dat236x-1 > accessed 14 February 2019.

[56] Domingos, supra, note 50.

[57] S Olhede and P Wolfe, "When algorithms go wrong, who is liable?" (2017) 14(6) Significance 8.

[58] S Barocas and AD Selbst, "Big data's disparate impact" (2016) 104 Cal L Rev 671.

[59] SC Olhede and PJ Wolfe, "The growing ubiquity of algorithms in society: implications, impacts and innovations" (2018) Trans R Soc A 1, 6.

[60] A Caliskan et al, "Semantics derived automatically from language corpora contain human-like biases" (2017) 356 Science 183.

[61] B Goodman and S Flaxman, "European Union regulations on algorithmic decision-making and a 'right to explanation'" (2017) 38(3) AI Magazine 1, 3.

[62] A Chouldechova, "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments" (2017) 5(2) Big Data 153.

are preferable candidates.[63] In short, machine learning can reify existing patterns of discrimination.[64]

Another flaw rooted in data is if foundational issues in measurement are neglected that apply in statistical research when using data for ML algorithms. Illustrative is the example of Google Flu Trends, which massively overestimated the prevalence of flu.[65] Lazer et al note that even with large quantities of data, one must still check validity, reliability and dependencies among data.[66] Traditional statistical methods could be used to extract relevant information that the tool overlooked, such as identifying the seasonality of the flu.

In short, algorithms may produce biased outcomes or decisions if the input data were biased, if relevant information was overlooked, and even if, in interpreting the outcomes, it is overlooked that biases in society are reflected in the input data.

b. Testing

Output may also be biased if the training data is not representative of the real-world environment in which the system is supposed to perform. An algorithm may produce unexpected and undesirable results if it encounters a situation that is markedly different from the data it was trained on. In a recent study, a computer program out performed human dermatologists at spotting skin cancer from photographs. A possible risk is that since the algorithm was trained using pictures of predominately white skin, it could perform worse for patients with darker skin.[67] When applying an algorithm and interpreting its outcome, it should always be verified if the algorithm is likely to produce a reliable result for these circumstances, given the environment it was tested in.

This raises the question of what a suitable testing regime might look like. A first relevant consideration in testing the algorithm is to select the type of algorithm. A suitable testing regime may test a variety of algorithms for their performance on the problem against the chosen performance measure, meaning how the solution to the problem is to be evaluated. Depending on how the predictions made by the algorithm are to be measured, a different algorithm may be appropriate.

Next, testing the algorithm involves selecting a test dataset and a training dataset. An algorithm will be trained on the training dataset and will be evaluated against the test set. Both datasets need to be representative of the problem that the algorithm is to solve. If the training dataset contains defects, it should be avoided that these defects are repeated in the testing dataset. A thorough testing regime tests the algorithm against realistic scenarios to ensure that it is fit for its purpose.

---

[63] J Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women" (*Reuters*, 10 October 2018) < www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G > accessed 14 February 2019.

[64] Goodman and Flaxman, supra, note 61, 3.

[65] D Butler, "When Google got flu wrong" (*Nature*, 13 February 2013) < www.nature.com/news/when-google-got-flu-wrong-1.12413 > accessed 14 February 2019.

[66] D Lazer et al, "The Parable of Google Flu: Traps in Big Data Analysis" (2014) 343 Science 1203, 1203, referring to D Boyd and K Crawford, "Critical Questions For Big Data" (2012) 15(5) Inform Commun Soc 662.

[67] KE Foley, "A pioneering computer scientist wants algorithms to be regulated like cars, banks, and drugs" (*Quartz*, 3 June 2017) < qz.com/998131/algorithms-should-be-regulated-for-safety-like-cars-banks-and-drugs-says-computer-scientist-ben-shneiderman/ > accessed 14 February 2019.

c. Decision models

As algorithms become more sophisticated, their decision-making process may become less tractable. The problems associated with biases in data and insufficient testing may get more pervasive as the decision model gets more complex, because these biases may be more difficult to predict or identify.

For instance, avoiding biased results rooted in social inequalities is difficult if sensitive information, such as ethnicity, is correlated with seemingly neutral variables, such as home address. In such cases, removing the sensitive variable will not prevent the biased result. Sophisticated algorithms may be able to reconstruct sensitive information from other inputs, even if they are not given this information.[68] An example was the Netflix algorithm that reportedly presented some users with racially targeted promotional images for films and TV shows, despite not asking members for information on their race.[69] With sufficiently large data sets, the task of exhaustively identifying and excluding data features correlated with "sensitive categories" a priori may be impossible. If we are not aware of correlations between variables, these hidden relationships may obscure the rationale for how predictions are being made.[70]

The chosen decision model may also turn out to be unsuitable if the real-world environment behaves differently from what was expected. If an algorithm is fed with input from users or consumers, designers may have to account for the interaction with the social-learning abilities in the decision model. Microsoft learned this lesson the hard way in 2017, when its chatting bot *Tay* had to be shut down after 16 hours because it became racist, denied the Holocaust and supported Hitler.[71] Tay was intended to be empathetic and was highly successful at that. However, this meant that as Twitter users started deluging Tay with racist, homophobic and otherwise offensive comments, Tay stopped being family-friendly and its type of language changed dramatically. Since then, Microsoft launched a new bot that was programmed differently with respect to its responses to input.[72]

As these examples illustrate, it will not always be possible to attribute the problem to one aspect of the algorithm. Nevertheless, these aspects – data, testing and decision model – can be useful starting points for identifying biases associated with algorithms.

IV. Is requiring transparency feasible and useful?

Understanding the roots of flaws and biases of algorithms allows us to evaluate the idea of a transparency requirement for algorithms. An effective transparency requirement would need to offer an explanation that is both feasible and useful. A feasible definition

[68] F Doshi-Velez et al, "Accountability of AI Under the Law: The Role of Explanation" (2017) arXiv:1711.01134, 8–9.

[69] L Bakare, "Netflix – treat your black customers with the respect they deserve" (*The Guardian*, 24 October 2018) < www.theguardian.com/commentisfree/2018/oct/24/netflix-black-customers-actors-hollywood-diversity > accessed 14 February 2019.

[70] Olhede and Wolfe, supra, note 57, 4; Goodman and Flaxman, supra, note 61, 4.

[71] Perez, supra, note 6.

[72] K Johnson, "Microsoft's Zo chatbot refuses to talk politics, unlike its scandal-prone cousin Tay" (Venture Boat 5 December 2016) < venturebeat.com/2016/12/05/microsofts-zo-chatbot-refuses-to-talk-politics-unlike-its-scandal-prone-cousin-tay/> accessed 14 February 2019.

of transparency allows programmers or algorithm producers to comply with the requirement. A useful definition of transparency provides sufficient information to plaintiffs, defendants and courts in legal cases involving algorithms.

1. Feasible transparency of algorithms

Transparency[73] means tracing back how certain factors were used to reach an outcome in a specific situation. Ideally, this means answering what were the main factors in the decision, how changing a certain factor would have changed the decision and, if applicable, what factor resulted in different decisions in two cases that look similar.[74] The concrete interpretation of transparency depends on the context in and purpose for which it is used.

When demanding explanation from humans, we typically want to know how certain input factors affected their final decision.[75] We may ask them about the main factors in the decision, how changing a certain factor would have changed the decision, and what factor resulted in different decisions in two similar cases. This is not to say that humans are always perfectly able to explain their decisions. We may rationalise our decisions when asked about them, listing socially correct reasons rather than our actual motivations.

With respect to algorithms, a motivation is limited to correlation in the statistical sense: it does not tell us why the outcome is as it is.[76] Given this limitation, three approaches to transparency could be distinguished.

First, a transparency requirement could focus on the input: the training, testing and operational data. Having access to this data could allow an observer to detect biases present in the dataset on which the algorithm operated or in society at large, as discussed above. Algorithms could be reviewed for their neutrality based on the accuracy and characteristics of the algorithm's inputs.

A second possibility is to require transparency of the decision-making process of the algorithm. Following this approach, an explanation should permit an observer to determine how input features relate to predictions, and how influential a particular input is on the output. This presumes that the model of the algorithm can be articulated and understood by a human.[77] An explanation of the decision model of the algorithm may become increasingly difficult as the algorithm becomes more complex. Even if an algorithm's source code were made transparent, it would only give a snapshot of its functionality. This is particularly true for adaptive, self-learning systems.[78]

[73] Some authors distinguish between transparency and explanation of algorithms' decisions (eg Doshi-Velez et al supra, note 68, 6). I use the terms interchangeably, focusing on explainability in concrete cases as a means to achieve transparency of algorithms overall.

[74] Doshi-Velez et al, supra, note 68, 3.

[75] Doshi-Velez et al, supra, note 68, 1; S Wachteret al, "Why a right to explanation of automated decisionmaking does not exist in the general data protection regulation" (2017) 7(2) International Data Privacy Law 76.

[76] M Hildebrandt, "Defining Profiling: A New Type of Knowledge?" in M Hildebrandt and S Gutwirth (eds), *Profiling the European Citizen* (Springer 2008).

[77] Goodman and Flaxman, supra, note 61, 6.

[78] M Ananny and K Crawford, "Seeing without knowing: limitations of the transparency ideal and its application to algorithmic accountability" (2016) 20(3) New Media Soc 973, 982.

Finally, transparency could be required in terms of the outcomes or decisions of the algorithm. In some cases it may be obvious that the outcome or decision reached by an algorithm is harmful. In others, analysing the outcomes may show harm, for instance in the form of (statistical) discrimination. This highlights the importance of counterfactuals for explanation of algorithmic decisions.[79] An explanation could be provided by probing the AI system with variations of the original inputs changing only the relevant variable, to see if the outcomes are different.[80]

The first and last approach to transparency may be more feasible for programmers than the approach focusing on the decision-model of the algorithm. As technology advances, more instruments may become available to quantify the degree of influence of input variables on algorithm outputs.[81] Research is also underway in pursuit of rendering algorithms more amenable to ex post and ex ante inspection.[82] Nonetheless, generating explanations of an algorithm is a non-trivial engineering task that takes time and effort that could also be spent on other goals.[83]

2. Useful transparency of algorithms

Depending on the circumstances, the need for transparency of algorithms and the type of information needed is likely to differ considerably. For instance, while one may be interested to know why Netflix' algorithm is suggesting certain content, one may have a more legitimate interest in getting an explanation for an algorithm that decides on loans, jobs or parole.

Depending on who needs transparency, an explanation may, moreover, take a different form. When defining a requirement for transparency we must consider who needs to understand the decision-making processes of the algorithm in question, and ask what purpose the understanding is to serve.[84] The justifications for requiring transparency determine what information should be provided about the algorithm and to whom it should be disclosed.

Let us consider two examples of possible purposes for understanding the decision-making process of an algorithm.

a. Civil liability cases

Transparency of algorithms may help courts to assign appropriate responsibility for failures in decision-making. To determine negligence liability for a decision based on an algorithm, courts would need to know how the algorithm reached its decision, or where it may have been flawed.

[79] S Wachter et al, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR" (2017) arXiv preprint arXiv:1711.00399.

[80] Doshi-Velez et al, supra, note 68, 7.

[81] See, for instance, A Datta et al, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems" (2016) Security and Privacy, 2016 IEEE Symposium, 598.

[82] See, for instance, R Jia and P Liang, "Data recombination for neural semantic parsing" (2016) Association for Computational Linguistics 12; A Vellido et al, "Making machine learning models interpretable" (2012) ESANN proceedings 163.

[83] Doshi-Velez et al, supra, note 68, 3.

[84] Reed, supra, note 26, 6.

In order for a negligence claim to prevail, three elements must normally be demonstrated: (i) the defendant had a duty of care; (ii) the defendant breached that duty; and (iii) that breach caused an injury to the plaintiff. Proving causation may be difficult when an algorithm is involved. If multiple parties are involved in the design and programming of the algorithm, it may be more difficult to establish a reasonable connection between the act or omission of the defendant and the damage.[85] Proving the connection between the act of the algorithm and the damage may be particularly difficult when the algorithm recommended an action (comparable to an expert) rather than taking an action (such as a self-driving car).[86] Related to this, courts may have difficulty determining whether the harm was foreseeable for the defendant, considering that unforeseeable behaviour is to some extent intended by the designers of ML algorithms and systems using neural networks.[87]

This means that output transparency of the algorithm may not be useful, in the sense that harmful output would still have to be attributed to the algorithm. In these circumstances, transparency of the input and the decision-model could help clarify what went wrong in case of harm. Any explanation of the decision-making process of the algorithm should inform courts of the factors influencing the decision and the possible biases in the system. Nevertheless, for this information to truly improve the accuracy of assigning legal responsibility, courts would need some technical literacy. With respect to arguing and evaluating how complex algorithms may have caused harm, it is likely that courts and injured parties remain at a disadvantage compared to the experts producing the algorithm.

Moreover, all the factors covered by a possible transparency requirement would need to be clear from the outset. Differently from humans, algorithms cannot be asked ex post, in a litigation scenario, to refine their explanations without additional training data.[88] Moreover, the relevant information would need to be explained in a manner that judges and parties to a case would understand. Differences in technical literacy mean that, for most people, simply having access to the underlying code of an algorithm is insufficient.[89]

In short, if a transparency requirement is to aid parties in liability cases, the demands are high. The explanation would need to put courts and parties in a position to evaluate whether an algorithm was biased or flawed, and whether this constitutes negligence on the part of the producers.

b. Breach of fundamental rights

A person involved in a potential breach of fundamental rights may need to know how an algorithm makes its decisions. One could consider a criminal case in which an algorithm

[85] ME Gerstner, "Comment, Liability Issues with Artificial Intelligence Software" (1993) 33 Santa Clara Law Review 239.

[86] JKC Kingston, "Artificial Intelligence and Legal Liability" in M Bramer and M Petridis (eds), *Research and Development in Intelligent Systems* XXXIII. SGAI (Springer, Cham 2016) 274.

[87] Scherer, supra, note 13, 366.

[88] Doshi-Velez et al, supra, note 68, 9.

[89] J Burrell, "How the machine 'thinks': understanding opacity in machine learning algorithms" (2016) 3(1) Big Data & Society 2053951715622512.

predicting a defendant's recidivism risk is used to determine that defendant's sentence.[90] Another example may be a hiring algorithm used to select candidates for jobs. In such cases, the harmed party would need the algorithm to be explained in order to be able to substantiate his claim that the algorithm may be biased.

While input transparency may reveal social inequalities or biases, it is not sufficient to know whether sensitive information was used as input variables for the algorithm. On the one hand, the algorithm may be able to create proxies to reconstruct the sensitive information from other inputs.[91] On the other, neutral data that can reveal sensitive information does not necessarily result in discriminatory outcomes.[92]

Input transparency combined with access to the decision model may allow harmed parties to probe the algorithm with variations of the original inputs, to check whether the outcomes are different. Another approach may be to reduce the legal evidentiary burden when algorithms are involved, allowing for output transparency to suffice. In some circumstances, such as employment cases in the US, statistical evidence that the outputs of a decision-making process disproportionately exclude a particular race or gender can be sufficient to shift the burden of proof on the decision-maker.[93] Such an approach to explanation or transparency of algorithms could be considerably more workable in practice than requiring details on how the sensitive information affected the decision in a concrete case – and thereby more useful.

Nonetheless, aligning the decision models of more advanced algorithms with the requirements of fundamental rights may remain difficult. The decision models of ML algorithms are based on past decisions and outcomes using complex optimisation models, whereas fundamental rights aim to protect the individual, using reasoning and interpretation.[94] We may find that no approach to transparency meets the standards of fundamental rights procedures. If so, we need to ask in which contexts we may prefer human over automated decision-making.

3. When should we require transparency of algorithms?

We could imagine myriad legal situations in which it could be useful that algorithms are more transparent. However, designing a system that generates useful explanations takes valuable time and engineering effort. Therefore, we need to ask when it is worth requiring transparency, balancing the utility of transparency against the costs of generating such a system.

The usefulness of transparency may depend on the risk associated with the decision. Regulatory transparency requirements should be context-dependent and based on risks to safety, fairness, and privacy.[95] We may require transparency for decisions that have a

[90] E Thadaney Israni, "When an Algorithm Helps Send You to Prison" (*NY Times*, 26 October 2017) <www.nytimes.com/2017/10/26/opinion/algorithm-compas-sentencing-bias.html> accessed 14 February 2019.

[91] Section III.2.c.

[92] T Calders and S Verwer, "Three naive Bayes approaches for discrimination-free classification" (2010) 21(2) Data Mining and Knowledge Discovery 277.

[93] Doshi-Velez et al, supra, note 67, 6; JD Cummins and B Isle, "Toward systemic equality: Reinvigorating a progressive application of the disparate impact doctrine" (2017) 43(1) Mitchell Hamline Law Review 102.

[94] Burrell, supra, note 89.

[95] S Wachter et al, "Transparent, explainable, and accountable AI for robotics" (2017) (6) Science Robotics.

great impact on someone besides the decision-maker. For less important decisions, we may opt for less transparency in favour of a better system performance.[96] In those settings, we may be able to rely on statistical measures of a system's performance to provide transparency, without providing an explanation for each given decision. Other factors determining the need for transparency may be whether there is value in knowing whether the decision was made erroneously, and whether there is reason to believe that an error has occurred.[97]

There may be a trade-off between the capacity of a model and its explainability. Easy to interpret linear models may only be able to explain simple relationships. Complex methods, representing a rich class of functions, may be hard to interpret.[98] By requiring more transparency, we may need to accept that systems are less accurate than they could technically be. This is relevant considering that both clarity and accuracy are useful for preventing errors.[99] We should also consider that the opacity of algorithms may have non-technical justifications, such as protecting trade secrets. Finally, we need to be aware of the costs of requiring transparency particularly for small companies, which may be disproportionally burdened. If the costs of providing transparency are high, requiring it can potentially have negative effects on innovation. Despite the potential burden on developers of requiring transparency, this may nevertheless be justified for algorithms involved in decisions which will have a great impact.

Regardless, for any transparency requirement we need to consider whether a technically feasible explanation would be useful for the prospective recipients. Merely bombarding people with information they cannot process will not help them to make better decisions. Neither will a requirement for producers of algorithms to hand over the code help courts in assigning liability. Improving the technical literacy of legislators and judges may be part of the solution for some types of cases. In other cases, we may need to accept that transparency is simply not feasible in practice. If so, it may be necessary to consider whether the algorithm presents such a great risk that limitations to its use are justified.

A more general question is what level of transparency we should require from algorithms. One could argue that it should be high, as we are not just trying to replace humans with equally flawed systems, but aiming to improve on them. At the same time, we need to keep in mind that, while biases in algorithms may be difficult to identify, humans may be unaware of their own biases as well. Over time, we may attain more understanding of these biases than we do now if algorithmic process can demonstrate the transparency of a decision made by either a human or a machine.[100]

Recent calls for transparency, as well as the requirements of the GDPR, do not yet specify what transparency would concretely entail in various contexts. We need to ask whether a technically feasible explanation would be useful for the prospective user, and

[96] M Oswald, "Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power" (2018) Phil Trans R Soc A 376, 20170359.

[97] Doshi-Velez et al, supra, note 68.

[98] Goodman and Flaxman, supra, note 61, 6.

[99] Doshi-Velez et al, supra, note 68, 10.

[100] C Kuner et al, "Machine learning with personal data: is data protection law smart enough to meet the challenge?" (2017) 7(1) Int Data Priv L 1, 2.

whether the value of the explanation is worth the costs of generating it. In some cases, we may find that access to data suffices for the purpose in question. In other contexts, it may not be technically feasible to produce any useful form of transparency. Before jumping to regulatory conclusions regarding the need for transparency of algorithms, or even of AI, we should give more thought to what this means in particular legal contexts.