UK Parliament POST

POSTNOTE

Number 633 October 2020

Interpretable machine learning

This POSTnote gives an overview of machine learning (ML) and its role in decision-making. It examines the challenges of understanding how a complex ML system has reached its output, and some of the technical approaches to making ML easier to interpret. It gives a brief overview of some of the proposed tools for making ML systems more accountable, such as algorithm audit and impact assessments.

Overview

- Machine learning (ML) is being used to support decision-making in applications such as recruitment and medical diagnoses.
- Concerns have been raised about some complex types of ML, where it is difficult to understand how a decision has been made.
- A further risk is the potential for ML systems to introduce or perpetuate biases.
- Approaches to improving the interpretability of ML include designing systems using simpler methods and using tools to gain an insight into how complex systems function.
- Interpretable ML can improve user trust and ML performance, however there are challenges such as commercial sensitivity.
- Proposed ways to improve ML accountability include auditing and impact assessments.

Background

Machine learning (ML), a type of artificial intelligence (AI, Box 1), is increasingly being used for a variety of applications from verifying a person's identity based on their voice to diagnosing disease. ML has the potential to bring many social and economic benefits, including increased labour productivity and improved services across a wide range of sectors.[1–3]

However, there are concerns that decisions that are made or informed by the outputs of ML can lack transparency and accountability. This can be a particular issue for certain types of ML (such as deep learning, Box 1), where in some cases it may not be possible to explain completely how its outputs have been produced. Furthermore, ML systems can be susceptible to introducing or perpetuating discriminatory bias (Box 2). Experts have warned that a lack of clarity on how ML decisions are made may make it unclear whether the systems are behaving fairly and reliably, and may be a barrier to wider ML adoption.[4,5]

In 2018, the Lords Committee on AI called for the development of AI systems that are "intelligible to developers, users and regulators". It recommended that an AI system that could have a substantial impact on an individual's life should not be used unless it can produce an explanation of its decisions.[4] In a January 2020 review, the Committee on Standards in Public Life (a public body) noted that explanations for decisions made using ML in the public sector are important for public accountability and recommended that government guidance on the public sector use of AI should be made easier to use.[6]

The UK Government has highlighted the importance of ethical ML,[7–9] and the risks of a lack of transparency in ML-assisted decision-making.[10] In 2018, it published a new version of its Data Ethics Framework, setting out guidance on how data should be used in the public sector.[11] It also established the Centre for Data Ethics and Innovation to provide independent advice on measures needed to ensure safe, ethical and innovative uses of AI.[10,12] The Information Commissioner's Office (ICO) and the Alan Turing Institute co-produced guidance in 2020 to support organisations in explaining AI-assisted decisions to individuals affected by them.[13]

Machine learning and algorithms

All AI systems are underpinned by an algorithm or a set of algorithms. An algorithm is a set of instructions used to perform tasks (such as calculations and data analysis), usually using a computer.[4] Traditional approaches to coding algorithms involved a large number of pre-programmed rules,[13] however, ML algorithms allow systems to learn using example data (referred to as 'training data'), without requiring all instructions to be explicitly programmed.[4,14] ML algorithms are not new, but their capability has significantly improved in recent years due to development of more sophisticated algorithms, greater availability of training data and advances in computing power.[2,8]

The Parliamentary Office of Science and Technology, Westminster, London SW1A 0AA
02072192840 post@parliament.uk post.parliament.uk @POST_UK

Box 1. Definitions

Artificial intelligence (AI)
There is no universally agreed definition of AI. It is defined in the Industrial Strategy as "technologies with the ability to perform tasks that would otherwise require human intelligence, such as visual perception, speech recognition, and language translation".[15] AI is useful for identifying patterns in large sets of data and making predictions.[16]

Machine learning (ML)
ML is a branch of AI that allows a system to learn and improve from examples without all its instructions being explicitly programmed.[2] An ML system is trained to carry out a task by analysing large amounts of training data and building a model that it can use to process future data, extrapolating its knowledge to unfamiliar situations.[2] Applications of ML include virtual assistants (such as Alexa), product recommendation systems, and facial recognition.[2] There is a range of ML techniques, but many experts attribute recent advances to developments in deep learning:

- **Artificial neural networks (ANNs).** Type of ML that have a design inspired by the way neurons transmit information in the human brain.[17] Multiple data processing units (nodes) are connected in layers, with the outputs of a previous layer used as inputs for the next.[18,19]
- **Deep learning (DL)**. Variation of ANNs. Uses a greater number of layers of artificial neurons to solve more difficult problems.[16] DL advances have improved areas such as voice and image recognition.[20]

Particular progress has been made in deep learning (a type of ML, Box 1).[2,20] ML relies on large datasets to train its underlying algorithms. In general, the more data used to train an ML system, the more accurate its outputs.[21] However, the quality of the data is also important; unrepresentative, inaccurate or incomplete data can lead to risks such as bias (Box 2).

ML in decision-making

Modern ML systems are increasingly used to inform decision-making in a variety of different applications, such as sifting recruitment candidates,[22] predicting criminal reoffending[23] and analysing medical data to help clinical diagnoses.[24] Box 3 describes legislation relevant to ML-assisted decision-making.

Human involvement

In some cases, ML is used in decision support, meaning that the output of the ML system is used to inform the thinking of a human decision-maker alongside other information available to them.[23,25] In other cases ML is used to automate decision-making; the output of an ML system and any action taken as a result (the decision) is implemented without human involvement.[26,27] Most experts agree that for some applications, retaining a degree of human oversight is important, particularly for applications that may have significant impacts on people.[6]

Black Box ML

Some types of ML, such as deep learning (Box 1) are very complex, meaning it may be difficult or impossible to fully understand how a decision has been reached.[28,29] These systems are commonly referred to as 'black box' ML.[30] This term is also used to describe ML systems whose workings are purposely concealed, for example because the technology is proprietary.[31] Academics and others have highlighted that a lack of transparency in how these systems function makes it difficult to verify their safety and reliability.[5,31] This has prompted a growing interest in the field of 'interpretable ML'.

Box 2: Algorithmic bias

The term 'algorithmic bias' is commonly used to describe discrimination against certain groups on the basis of an ML system's outputs.[32–34] Some high-profile examples include:

- A 2019 study of an algorithm used to allocate healthcare in US hospitals found that it was less likely to refer Black people than White people who were equally sick to healthcare programmes.[35,36]
- A 2015 study found that when a web user's gender was set to female, Google's online advertising system showed fewer high-income job adverts than it did to male users.[37]

Bias can be introduced into an ML system in different ways, including:[38]

- **Training data.**[38] Insufficient training data about certain demographics can lead to ML algorithms being less accurate for those groups.[38–42] For example, research shows that if a facial recognition algorithm is trained solely on the faces of White people, it performs more accurately for that group than others.[41] Algorithms can also reflect historic biases that exist in the data they are trained on.[32,43] For example, there have been widespread concerns about the potential for ML systems to exhibit racially biased outcomes as a result of being trained on historic crime data that contains racial discrimination.[44–47]
- **Decisions made in development.** ML developers make decisions and assumptions at various stages of a system's development,[48] including what attributes they want an algorithm to consider, how data will be categorised, how the ML system is optimised and what training data are used.[38,49–51] These may result in a model that has inadvertent discriminatory features.[52] Some stakeholders have suggested that a lack of diversity in ML research and development teams could contribute to this issue.[39,52,53]

There are also broader risks around the way humans interact with ML outputs, which could lead to poor or unfair outcomes when ML is used in decision-making.[32] In some cases, individuals may become over-reliant on an ML system and follow its advice without considering other factors or applying their own knowledge and experience.[32,54] Conversely, there is also a risk of 'algorithm aversion', where users may not accept or trust an ML output even when the system performs well at a task.[55–59]

Making ML interpretable

The term 'interpretability' is typically used to describe the ability to present or explain an ML system's decision-making process in terms that can be understood by humans (including AI developers, users, procurers, regulators and decision recipients).[60–63] Terminology in this area varies and is inconsistent (other common terms include 'explainability' and 'intelligibility').[2,4,8,60,64–66] Many stakeholders have highlighted that the extent to which ML needs to be interpretable is dependent on the audience and context in which it is used.[5,13,67] Some have emphasised that there is no 'one-size-fits-all' approach to interpretable ML and consideration should be given to what information an individual may require and why.[13,68]

Technical approaches to interpretable ML

Technical approaches to interpretable ML include designing systems using types of ML that are inherently easy to understand and using retrospective tools to probe complex ML systems to obtain a simplified overview of how they function.[5]

Box 3: Legal framework for ML decision-making

There is no UK regulation specific to ML. However, there is a range of existing legal frameworks that apply to its use. For example, UK data protection law has specific provisions around automated decision-making. Human rights law and administrative law may also apply to certain ML applications.[25,69,70] ML-based decisions must also comply with the Equality Act 2010, which prohibits certain kinds of discrimination based on protected characteristics.[71] In addition, there may be relevant sector-specific regulations. For example, some ML-based healthcare apps and software must comply with medical directives provided by the Medicines and Healthcare products Regulatory Agency.[72]

Data protection law

The Data Protection Act 2018 and the EU GDPR regulate the collection and use of personal data.[13] If ML uses personal data (including in its training and deployment), it falls under this legislation.[13] GDPR prohibits fully automated decision-making that has a "legal or similarly significant effect" on a person, unless that person has given consent, it is necessary to fulfil a contract, or it is required by law.[73–75] In cases where a fully automated decision has such an effect, individuals must be informed of its existence, provided with meaningful information about it, able to challenge the decision and able to obtain human intervention.[13,74–76]

Right to an explanation

The extent to which GDPR provides the right for individuals to receive an explanation of an automated decision made about them is an area of debate.[13,77–80] GDPR interpretive guidance (the 'recitals') states that an individual has a right to receive an explanation of an automated decision. However, some academics have raised concerns that this is not enforceable in practice, as the recitals are not legally binding.[32,77,81] The ICO has stated that an individual must be given an explanation to enable their right to receive meaningful information about a decision.[13] Experts have also raised concerns about GDPR being limited to fully automated decision-making as, in practice, few significant decisions are fully automated.[39,77–79] Some have called for the law to be strengthened to include the right to an explanation in cases where ML is part of a decision.[23,77,82]

Using interpretable ML models

Some types of ML are interpretable by design, meaning that their complexity is restricted in order to allow a human user to understand how they work.[5,13,60] However, some ML applications, such as identifying anomalies in video footage, may rely on the use of black box ML techniques, including deep learning. Substituting for an easier to understand type of ML may be difficult for such applications, as it may not be possible to achieve the same level of performance accuracy.[5,21,83] Some stakeholders have said that limiting applications to interpretable techniques would, in some cases, limit the capability of ML technology.[60,84,85] However, others argue that there is not always a trade-off between accuracy and interpretability and that in many cases complex ML can be substituted for a more interpretable method.[31] Another approach that has been proposed is to use a 'decomposable' ML system, where the ML's analysis is structured in stages and interpretability is prioritised for the steps that most influence the output.[62]

Some stakeholders have said that ML that is not inherently interpretable should not be used in applications that could have a significant impact on an individual's life (for example, in criminal justice decisions).[4,31,86] The ICO and Alan Turing Institute have recommended that organisations prioritise using systems that use interpretable ML methods if possible, particularly for applications that have a potentially high impact on a person or are safety critical.[83]

Tools for interpreting black box ML

An active area of ML research seeks to develop techniques to understand how complex ML systems function internally, and to help explain their outputs.[4,5,60,66,87] As some of these techniques are in early stages of development, their use is not currently widespread. Some tools aim to interpret a specific ML decision, while others can be used to give a broad understanding of how an ML system behaves. The tools also vary in the information they provide. Some highlight the features of the input data that contributed most to the outcome, while others provide a simplified overview of the ML system. Examples include:

- **Proxy models** (or surrogate models) provide a simplified version of a complex ML system.[88] They are created by testing how a complex ML algorithm responds to different input data and building a model that approximately matches it.[89] Proxy models can provide useful insights into the behaviour of a more complex model, but may be limited in how fully they represent its behaviour.[31] There is also a class of related techniques that can be used to approximate how an ML system arrives at an individual decision.[88]
- **Saliency mapping or visualisation** is particularly useful for understanding why an image classification algorithm has classified an image in a certain way.[90,91] It works by creating a visual map highlighting the parts of the image (or other data) that most influenced the ML system's output.[92]
- **Counterfactual explanations** aim to illustrate the changes in input data that would be needed in order to get a different outcome from an ML system, and can be used to explain an algorithm's output in individual cases.[93] For example, a counterfactual explanation for a system that has rejected a loan application would tell a user what changes would be needed in the input data (such as a person's income) in order to have the application approved.[93]

Some companies have developed open source tools based on some of these methods to help developers design more interpretable ML. For example, Microsoft and IBM have software toolkits to support ML developers.[94–96]

Explaining ML outcomes to individuals

The importance and purpose of explaining an ML decision to an individual – and the type of explanation that may be most useful – differs depending on the context in which ML is used.[5,97] For example, a citizens' jury of 36 participants run by the ICO found they thought that having an explanation of an ML-based decision was more important in job recruitment and criminal justice scenarios (such as selecting offenders for a rehabilitation programme), than in healthcare scenarios (such as stroke diagnosis).[97] In healthcare scenarios, participants valued the accuracy of the decision more than having an explanation of it.[5] Not all ML systems require an explanation of their outcomes, as they may not have a significant impact on an individual (for example, a product recommender system).[98]

The ICO and Alan Turing Institute have produced guidance for organisations to help them explain AI-based decisions.[83] It

includes guidance on selecting the type of ML to use, tools for interpreting complex ML (such as those in the previous section) and how a system's outputs can be explained to an individual.

Benefits and challenges to interpretability

Interpretable ML can have potential benefits for organisations, individuals and wider society, including:

- **Improved performance.** Greater interpretability can give developers a better understanding of how an ML system functions and how to improve it.[5,13] It can help to verify that the system is performing safely and robustly, identify flaws, and ensure potential biases (Box 2) are mitigated.[5,99]
- **Improved user trust.** Some research has found that explaining how a system has reached its outcome can increase public trust and confidence in ML. However, the relationship between explanations of ML and user trust is complex and depends on factors such as the ML application and type of explanation given.[100] In some cases, there is a risk that explanations of ML can be misleading (see below).
- **Regulatory compliance.** Using interpretable ML or explaining decisions to individuals may help ensure that its use complies with relevant legislation (Box 3).[13,101]

There are also reasons why organisations may not want to make certain information about their ML systems publicly available, and wider challenges with interpretable ML, including:

- **Commercial sensitivity**. Many companies regard their ML algorithms as valuable intellectual property and may be reluctant to give away information about how their programs work in case they lose their commercial advantage.[13,102]
- **Risk of gaming**. If users have access to information about how an algorithm works, it may be possible for them to manipulate the ML system (referred to as 'gaming').[5,13] For example, a fraud detection algorithm may look for certain traits in financial data that indicate fraud; if these are known, individuals may change their behaviour to avoid detection.[5]
- **Cost.** Organisations deploying ML may not prioritise explaining how their systems work as there may be costs and resources associated with doing so.[97]
- **Mistrust or deception**. In some cases, there are limitations to the reliability of ML interpretability tools. There is a risk that some tools may oversimplify how the ML works or give incomplete information about it, which may harm a user's trust.[103,104] There is also a risk that oversimplified explanations could cause users to develop a false sense of understanding of a system and over-rely on its outputs.[105,106]

Wider ML accountability tools

In addition to technical approaches to interpretable ML, many stakeholders have called for wider accountability mechanisms to ensure that ML systems are designed and deployed in an ethical and responsible way.[107,108]

Open and documented processes

Some stakeholders have suggested that algorithm developers should be required to produce detailed records about the algorithm, including documentation of its programming, training data and decisions made during devleopment.[109] This could make it easier for problematic decisions made by an ML system to be traced back.[109] The Partnership on AI has a project that aims to formalise guidance on such documentation.[110]

Machine learning fact sheets

Researchers have proposed the idea of fact sheets for ML systems and their underlying data.[111] A fact sheet could give consumers information about a system's characteristics, such as its performance, safety, security and limitations.[111–113] Fact sheets could also accompany datasets that are used to train ML, so that developers have a better idea of how the resulting system is expected to perform in different contexts (for example, in different population groups or locations).[114,115]

Algorithmic impact assessments

Algorithmic impact assessments (AIAs) have been proposed as a way for creators or procurers of algorithms to evaluate the impacts and potential risks of an ML system before deployment.[116] They can also be ongoing processes that organisations use to continually monitor their algorithms.[117] There is little consensus on what AIAs should involve and how they would be implemented.[117] Discussion about AIAs has mainly focused on their use in the public sector.[7,8] The AI Now Institute has proposed that public sector bodies should carry out AIAs and has developed an AIA framework.[116] In 2020, Canada made it mandatory for all public sector automated decision-making systems to undergo an AIA.[118]

Algorithm audit and certification

There are different definitions of algorithm audit and different proposals for what they could involve.[117,119] In many cases it refers to a type of regulatory inspection, whereby an algorithm is inspected by an external organisation (such as a government associated body) to ensure it complies with certain regulations or principles.[120–122] Audits can vary, but may involve examining an ML system's code, training data, process used to build it and outputs. While there have been growing calls for algorithm audits to take place,[4,123–125] there is currently no mandatory requirement for them in the UK. The ICO recently published an auditing framework to inform its investigation of an AI system's compliance with data protection law. Some have suggested that certification could be used to signify algorithms that have been audited, or to verify that they meet certain design standards.[126]

Principles, frameworks and standards

There are multiple examples of principles and codes of practice for ML (and AI more broadly), produced by public bodies and private sector organisations. Examples include those published by the Alan Turing Institute and UK Government,[7,8] the Lords AI Committee,[4] and tech companies.[127,128] A 2019 analysis found that there were 84 sets of ethical principles or guidelines for AI published globally.[129] Despite these initiatives, some stakeholders have expressed concerns that existing frameworks lack specific actions and coordination.[130–132] The Committee on Standards in Public Life notes that public bodies may be uncertain about which principles to follow and has recommended that guidance is made easier to understand.[6]

Several national and international bodies have started to produce industry standards to promote the ethical development of ML. In 2016, the British Standards Institution published the first UK standards for ethical design of robots and autonomous systems.[133] The Institute of Electrical and Electronics Engineers (a global standards body) is also working AI standards.[134]

Endnotes

1. Financial Conduct Authority (2019). Machine learning in UK financial services.
2. Royal Society (2017). Machine learning: the power and promise of computers that learn by example.
3. PwC (2017). AI analysis: sizing the prize.
4. House of Lords Select Committee on AI (2018). AI in the UK: ready, willing and able.
5. The Royal Society (2020). Explainable AI: the basics.
6. Commitee on Standards in Public Life (2020). Artificial Intelligence and Public Standards.
7. Office for AI and Government Digital Service (2019). Understanding artificial intelligence ethics and safety.
8. Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector.
9. HM Government (2018). AI Sector Deal.
10. Department for Digital, Culture, Media & Sport (2018). Centre for Data Ethics and Innovation Consultation.
11. Data Ethics Framework. *GOV.UK*.
12. HM Government (2019). AI sector deal one year on.
13. ICO and Alan Turing Institute (2020). Explaining decisions made with AI.
14. Points, L. *et al.* (2017). Artificial Intelligence and Automation in the UK.
15. UK Government (2017). Industrial Strategy: building a Britain fit for the future.
16. GO Science (2016). Artificial intelligence: opportunities and implications for the future of decision making.
17. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, Vol 61, 85–117.
18. Krogh, A. (2008). What are artificial neural networks? *Nature Biotechnology*, Vol 26, 195–197.
19. DSTL (2019). The Dstl Biscuit Book.
20. Jordan, M. I. *et al.* (2015). Machine learning: Trends, perspectives, and prospects. *Science*, Vol 349, 255–260.
21. Information Commissioner's Office (2020). Trade-offs. ICO.
22. Hymas, C. (2019). AI used for first time in job interviews in UK to find best applicants. *The Telegraph*.
23. The Law Society of England and Wales (2019). Algorithms in the Criminal Justice System.
24. Lysaght, T. *et al.* (2019). AI-Assisted Decision-making in Healthcare. *Asian Bioethics Review*, Vol 11, 299–314.
25. Oswald, M. (2018). Algorithm-assisted decision-making in the public sector: framing the issues using administrative law rules governing discretionary power. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, Vol 376,
26. Jefferies, D. (2019). What AI can do for the insurance industry. *Raconteur*.
27. CDEI (2019). Snapshot Paper - AI and Personal Insurance.
28. The Lancet (2018). Opening the black box of machine learning. *The Lancet Respiratory Medicine*, Vol 6, 801.
29. Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News*, Vol 538,
30. Head of maize Content What is the AI Black Box Problem? *maize*.
31. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, Vol 1, 206–215.
32. Centre for Data Ethics and Innovation (2019). Bias in Algorithmic Decision Making.
33. Wachter, S. (2019). Affinity Profiling and Discrimination by Association in Online Behavioural Advertising. *Berkeley Technology Law Journal*, Vol 35,
34. Koene, A. (2017). Algorithmic Bias: Addressing Growing Concerns. *IEEE Technology and Society Magazine*, Vol 36, 31–32.
35. Obermeyer, Z. *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, Vol 366, 447–453.
36. Ledford, H. (2019). Millions of black people affected by racial bias in health-care algorithms. *Nature*, Vol 574, 608–609.
37. Datta, A. *et al.* (2015). Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies*, Vol 1, 92–112.
38. Hao, K. (2019). This is how AI bias really happens—and why it's so hard to fix. *MIT Technology Review*.
39. House of Commons Science and Technology Committee (2018). Algorithms in decision-making.
40. Buolamwini, J. *et al.* (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research*, Vol 81, 1–15.
41. Klare, B. F. *et al.* (2012). Face Recognition Performance: Role of Demographic Information. *IEEE Transactions on Information Forensics and Security*, Vol 7, 1789–1801.
42. Grother, P. *et al.* (2019). Face recognition vendor test part 3:: demographic effects. National Institute of Standards and Technology.
43. Terzis, P. *et al.* (2019). Shaping the State of Machine Learning Algorithms within Policing: Workshop Report.
44. Lum, K. *et al.* (2016). To predict and serve? *Significance*, Vol 13, 14–19.
45. Ensign, D. *et al.* (2018). Runaway Feedback Loops in Predictive Policing. *Proceedings of Machine Learning Research*, Vol 81, 1–12.
46. Richardson, R. *et al.* (2019). Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice.
47. Angwin, J. *et al.* (2016). Machine Bias. *ProPublica*.
48. Kemper, J. *et al.* (2018). Transparent to whom? No algorithmic accountability without a critical audience. *Information, Communication & Society*, 2081–2096.
49. Danks, D. *et al.* (2017). Algorithmic Bias in Autonomous Systems. in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. 4691–4697. International Joint Conferences on Artificial Intelligence Organization.
50. Mittelstadt, B. D. *et al.* (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, Vol 3, 1–21.
51. Heilweil, R. (2020). Why algorithms can be racist and sexist. *Vox*.
52. World Economic Forum WEF white paper how to prevent discriminatory outcomes in machine learning.
53. West, S. M. *et al.* (2019). Discriminating systems. AI Now Institute.
54. Zerilli, J. *et al.* (2019). Algorithmic Decision-Making and the Control Problem. *Minds & Machines*, Vol 29, 555–578.
55. Onkal, D. *et al.* (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making*, Vol 22, 390–409.
56. Povyakalo, A. A. *et al.* (2013). How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography. *Med Decis Making*, Vol 33, 98–107.
57. Burton, J. W. *et al.* (2019). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, Vol 33, 220–239.

58. Dietvorst, B. J. *et al.* (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J Exp Psychol Gen*, Vol 144, 114–126.
59. Professor Nigel Harvey (2018). Written Evidence to Commons Science and Committee Committee inquiry into algorithms in decision-making.
60. Molnar, C. (2020). Interpretable Machine Learning.
61. Carvalho, D. V. *et al.* (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, Vol 8, 832. Multidisciplinary Digital Publishing Institute.
62. Lipton, Z. C. (2016). The Mythos of Model Interpretability. *ICML Workshop on Human Interpretability in Machine Learning*,
63. Hall, P. *et al.* (2019). An Introduction to Machine Learning Interpretability, Second Edition. O'Reilly.
64. Weller, A. Transparency: motivations and challenges. in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 23–40.
65. Mittelstadt, B. *et al.* (2019). Explaining Explanations in AI. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 279–288.
66. Arrieta, A. B. *et al.* (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion*, Vol 58, 82–115.
67. Bhatt, U. *et al.* (2020). Explainable machine learning in deployment. *FAT* 20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 648–657.
68. Wortman Vaughan, J. *et al.* (2020). A Human-Centered Agenda for Intelligible Machine Learning. Microsoft Research Lab.
69. Oswald, M. *et al.* (2018). Algorithmic risk assessment policing models: lessons from the Durham HART model and 'Experimental' proportionality. *Information & Communications Technology Law*, Vol 27, 223–250. Routledge.
70. Cobbe, J. (2019). Administrative law and the machines of government: judicial review of automated public-sector decision- making. *Legal Studies*, Vol 39, 636–655.
71. (2010). Equality Act 2010. Statute Law Database.
72. Medicines and Healthcare products Regulatory Agency (2020). Guidance: Medical device stand-alone software including apps (including IVDMDs).
73. (2019). Rights related to automated decision making including profiling.
74. European Commission General Data Protection Regulation (2016).
75. ICO (2019). Guide to the General Data Protection Regulation (GDPR).
76. UK (2018). Data Protection Act.
77. Wachter, S. *et al.* (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, Vol 7, 76–99.
78. Edwards, L. *et al.* (2017). Slave to the Algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law & Technology Review*, Vol 16, 18–84.
79. Edwards, L. *et al.* (2018). Enslaving the Algorithm: From a 'Right to an Explanation' to a 'Right to Better Decisions'? *IEEE Security & Privacy*, Vol 16, 46–54.
80. Ruiz, J. *et al.* (2018). Debates, awareness, and projects about GDPR and data protection. Open Rights Group.
81. Selbst, A. D. *et al.* (2017). Meaningful information and the right to an explanation. *International Data Privacy Law*, Vol 7, 233–242.
82. Big Brother Watch (2019). Big Brother Watch - submission to the CDEI Bias in Algorithmic Decision Making review.
83. ICO Explaining AI decisions part 2.
84. Ribeiro, M. T. *et al.* (2016). Model-Agnostic Interpretability of Machine Learning. *ICML Workshop on Human Interpretability in Machine Learning*,
85. Rane, S. (2019). The balance: Accuracy vs. Interpretability. *Data Science Ninja*.
86. AI Now (2017). AI now 2017 report.
87. AI Explanations Whitepaper. Google.
88. Ribeiro, M. T. *et al.* (2016). 'Why Should I Trust You?': Explaining the Predictions of Any Classifier. *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
89. Two Sigma (2019). Interpretability Methods in Machine Learning: A Brief Survey. *Two Sigma*.
90. Gilpin, L. H. *et al.* (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89.
91. (2017). Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer.
92. Simonyan, K. *et al.* (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *Workshop at International Conference on Learning Representations*,
93. Wachter, S. *et al.* (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, Vol 31, 842–887.
94. GitHub interpretml/interpret. *GitHub*.
95. Nori, H. *et al.* (2019). InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv*, Vol abs/1909.09223,
96. IBM AI Explainability 360.
97. ICO (2019). Project explAIn - interim report.
98. Doshi-Velez, F. *et al.* (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*,
99. The Challenges and Opportunities of Explainable AI. *Intel*.
100. (2019). Human-AI Collaboration Trust Literature Review - Key Insights and Bibliography. *The Partnership on AI*.
101. PricewaterhouseCoopers Explainable AI. *PwC*.
102. PricewaterhouseCoopers Opening AI's black box will become a priority. *PwC*.
103. Bhatt, U. *et al.* (2020). Machine Learning Explainability for External Stakeholders. *arXiv:2007.05408 [cs]*,
104. Papenmeier, A. *et al.* (2019). How model accuracy and explanation fidelity influence user trust. *arXiv:1907.12652 [cs]*,
105. Poursabzi-Sangdeh, F. *et al.* (2019). Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*,
106. Kaur, H. *et al.* (2020). Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14. ACM.
107. BBVA (2018). How to make artificial intelligence more ethical and transparent. *NEWS BBVA*.
108. ICO Big data, artificial intelligence, machine learning and data protection.
109. EU Commission (2020). White paper on artificial intelligence.
110. About us. *The Partnership on AI*.
111. Arnold, M. *et al.* (2019). FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity.

IBM Journal of Research and Development, Vol 63, 6:1-6:13.
112. Mojsilovic, A. (2018). Factsheets for AI Services: Building Trusted AI - IBM Research. *IBM Research Blog*.
113. IBM Research AI FactSheets 360.
114. Mitchell, M. *et al.* (2019). Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, 220–229.
115. Gebru, T. *et al.* (2020). Datasheets for Datasets. *arXiv*, Vol 1803.09010,
116. Reisman, D. *et al.* (2018). Algorithmic Impact Assessments. AI Now.
117. DataKind & Ada Lovelace Institute (2020). Examining-the Black Box.
118. Government of Canada (2020). Algorithmic Impact Assessment - Évaluation de l'Incidence Algorithmique.
119. Raji, I. D. *et al.* (2020). Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. *FAT* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44.
120. Tutt, A. (2017). An FDA for Algorithms. *Administrative Law Review*, Vol 69, 83–123.
121. Copeland, E. (2018). 10 principles for public sector use of algorithmic decision making. *nesta*.
122. Etzioni, O. *et al.* (2019). High-Stakes AI Decisions Need to Be Automatically Audited. *Wired*.
123. (2019). Report on Algorithmic Risk Assessment Tools in the U.S. Criminal Justice System.
124. O'Neil, C. (2018). Audit the algorithms that are ruling our lives. *Financial Times*.
125. Brundage, M. *et al.* (2020). Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv*, Vol 2004.07213,
126. IEEE-SA (2020). IEEE Invites Companies, Governments and Other Stakeholders Globally to Expand on Ethics Certification Program for Autonomous and Intelligent Systems (ECPAIS) Work.
127. Responsible AI principles from Microsoft. *Microsoft*.
128. Responsible AI Practices. *Google AI*.
129. Jobin, A. *et al.* (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, Vol 1, 389–399. Nature Publishing Group.
130. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, Vol 1, 501–507.
131. Morley, J. *et al.* (2019). From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Science and Engineering Ethics*, Vol 26, 2141–2168.
132. Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds & Machines*, Vol 30, 99–120.
133. BSI (2019). April 2019 - BSI's activities on Artificial Intelligence (AI).
134. IEEE (2019). Ethically Aligned Design.