# Accuracy and Explainability in Artificial Intelligence: Unpacking the Terms

*Short Paper*

**Dr Kathy McGrath**
Brunel University London
Uxbridge, Middlesex, UK
Kathy.McGrath@brunel.ac.uk

## Abstract

*Artificial intelligence (AI) has permeated many aspects of human life from product recommendations on retailers' websites to critical decisions affecting healthcare and law enforcement. As such systems become prevalent in high risk areas, explaining their logic to demonstrate issues such as fairness acquire increasing significance. Yet the current focus of machine learning models is the accuracy of decisions rather than their explainability. This paper analyses the findings from two citizens' juries convened to investigate the perceived trade-off between AI explainability and AI accuracy. While the official juries' report shows clear preferences for accuracy over explanation in some settings, this paper presents an alternative perspective informed by the concept of ambivalence. By introducing some additional metrics and highlighting the possibilities for different forms of explanation, this research demonstrates how the findings from the citizens' juries might be otherwise, and the social consequences arising. The paper concludes with some future research directions.*

**Keywords:** Artificial intelligence, explainability, interpretability, AI accuracy, citizens' jury, ambivalence

## Introduction

Research on artificial intelligence (AI) has been underway for more than half a century with mixed results, but in the last decade work in the sub-field of machine learning (ML) has addressed 'black box' decision support systems in which the program logic is not transparent to the user (Guidotti et al. 2018). These systems may be applied in areas like healthcare, advertising, finance, the automotive industry and the military to make decisions about such matters as patient referrals, consumer product preferences, bank loans, driver assistance, and security threats (e.g. Lipton 2018). Positive orientations towards such systems are based on their alleged capability to match human performance in decision making, while making more effective and cost-efficient use of resources. Proponents place particular emphasis on the accuracy of these ML models (London 2019). Negative sentiment focuses on practical and ethical concerns, such as trust, social inclusion and safety, given that little or no explanation of the decisions is available to those affected by them. Such sentiment emphasizes the issues arising because of the lack of transparency in the decision-making logic (Mendling et al. 2018). In short, a debate exists in the field of AI as to the relative importance of accuracy and explainability, arising from the perception amongst some researchers that achieving predictive accuracy in certain types of AI system, notably ML models, involves trading off some capability to explain the decisions made by the system. This research engages with this debate through a critical reinterpretation of empirical data collected from two citizens' juries convened to make recommendations to inform policy making about the types of AI system that would be appropriate in particular social contexts. The preliminary insights presented in this paper are part of an ongoing research project which asks the following questions:

- What are the social consequences of the accuracy versus explainability focus in AI developments;
- How might these developments affect the subjects of automated decisions?

The rest of this paper is organized as follows. The next two sections unpack the meanings of accuracy and explainability in AI, by introducing some additional metrics and highlighting different forms of providing an explanation. The following section describes the accuracy versus explainability debate in the field. This is followed by a discussion and reinterpretation of some key findings from two citizens' juries convened in the UK which were designed to explore whether individuals should receive an explanation when automated decisions are made about them, even if that impacts the performance of the AI system in terms of its accuracy. The paper concludes with some future research directions.

## Accuracy of AI Systems

The accuracy of AI systems is regularly reported by software vendors, researchers and the media publicizing the capability of such systems to match human decision-making performance in a highly cost-efficient way. The collaboration between Moorfields Eye Hospital in London and Google Health (formerly DeepMind) is a recent, well-publicized example of the use of machine learning in ophthalmology. The technology was applied to thousands of historic de-personalized scans to identify signs of eye disease and enable patient referral recommendations. Early results (De Fauw et al. 2018) described how machine learning technology matched the decision-making performance of world-leading experts for over fifty eye diseases, achieving 94% accuracy (Guardian 2018). But what is accuracy, and is it the key measure of system performance?

In AI, accuracy is one of four key metrics, the others being precision, recall (or sensitivity) and F1 score (Ghoneim 2019). When assessing the performance of a system, it is important to consider all four measures and prioritize them dependent on context. To illustrate the point, consider the four possible outcomes in the above eye disease example: correct/incorrect positive diagnosis (true/false positive – TP/FP) and correct/incorrect negative diagnosis (true/false negative – TN/FN). The formulae for the four measures are then (ibid.):

Accuracy = TP + TN / TP + TN + FP + FN

Precision = TP / TP + FP

Recall = TP / TP + FN

F1 Score = 2 x (Precision x Recall) / Precision + Recall

In effect, accuracy is the ratio of correct diagnoses to the total number of diagnoses, and while it is the measure most familiar to the general public, it is frequently neither the most meaningful nor the most useful. On the other hand, precision is the ratio of correct positive diagnoses to predicted positive diagnoses; recall is the ratio of correct positive diagnoses to actual positives in the dataset; and F1 score is the weighted average of precision and recall. Therefore, precision is a useful measure when there is a high cost associated with false positives, for example, wrongly identifying someone as a criminal or terrorist, whereas recall is useful when there is a high cost associated with a false negative, such as an undiagnosed cancer. The values of these measures can vary significantly. For example, in a dataset containing 1 million items, where 1,000 are true positives, a claim to 99% accuracy for a ML algorithm would sound truly impressive, yet the respective figures for precision and F1 score would be just 9% and 16.5% respectively.

Current thinking suggests that the prevalent machine learning approach of seeking to minimize errors so that machine-made decisions match human performance does not offer a sufficient basis for trust in such systems. For example, matched predictions may be made, but the training data for the model may include bias on the basis of race, ethnicity, or a wide range of social factors (Mendling et al. 2018). The ML model may be expected to learn such bias. Furthermore, predictions may be matched by percentage accuracy, but the cases in the matched samples may be different. In a clinical setting, patient safety could be at risk. In dynamic environments, such as consumer marketing, the basis of decision making may change frequently, but suitable training data for the model may not be available (Lipton 2018). Thus, models might be expected to perform sub-optimally, given that they lack the more human capabilities related to the transfer of prior learning to unfamiliar situations. Indeed, even if suitable training data is available, there would be a need for continuous improvement to train the model on the new data and test it rigorously (London 2019).

## Explainability of AI Systems

The explainability of machine learning models is a significant and complex issue, made more problematic because researchers disagree about what constitutes an appropriate explanation. This body of work is also inconsistent in its use of terminology, with some authors referring to explainability as a generic concept – meaning: capable of being explained – while others attach a very specific meaning concerned with revealing the detailed logic of a ML model. In the second case, the logic may be revealed through the creation of a post hoc model, such as a natural language explanation or visualization, which aims to explain the decision making of the initial black box model (Lipton 2018). Other researchers use the terms explainability and interpretability interchangeably (ibid.). In this paper, the term explainability is used in a generic sense, with references made to specific types of explanation where appropriate. Explainability is also viewed differently from *interpretability*, which refers to the extent to which a user can make sense of a ML model without needing to know details of its inner workings, and from *transparency*, which is the extent to which such a model is understandable to humans (Rudin 2019). Given these definitions, a fully interpretable model will be transparent so that how it works, for example, in terms of its training algorithm, individual components, or overall functioning will be understandable to a user. Crucially, though, forms of understanding and explanation can be achieved without the need for a full knowledge of the detailed workings of the model.

The capability to explain automated decision making is important for several reasons, including providing transparency and accountability for decisions made; supporting the management of risk, particularly in high stakes settings; and ensuring regulatory compliance. Unexplainable models also offer limited potential to uncover causal structure in observational data, which is an important concern in many application domains, not least in the prevention and treatment of diseases (Shortliffe and Sepulveda 2018). Where decisions cannot be explained, they also deprive people of necessary information to inform future behaviour, such as how to improve one's health or to qualify for a loan. From a regulatory perspective, machine learning models need to provide information as well as minimize errors. In light of the General Data Protection Regulation (GDPR) which became part of European Union law in May 2018 (EU 2016), individuals now have a right to 'meaningful information about the logic involved, as well as the significance and the envisaged consequences' (p. 42) when automated decisions are made about them. Furthermore, such decisions should be seen as fair and ethical (Faraj et al. 2018), which presents problems if the decision making logic cannot be explained.

## Accuracy vs Explainability in Deep Learning Systems

Evidently, there are diverse views about what constitutes accuracy and explainability in AI systems. The situation becomes even less clear in the specific case of deep learning systems, where the black-boxed nature of the models' logic is seen to be capable of little or no explanation while being highly accurate, creating the perception of a trade-off between these two system properties. Again, researchers are divided about how to respond to this situation and, in some cases, about whether such a trade-off is required now or in the future.

One school of thought argues that the 'ability to explain how results are produced can be less important than the ability to produce such results and empirically verify their accuracy' (London 2019, p. 15). Showing a preference for accuracy over explainability, London advocates rigorous testing of AI systems and use of regulation to limit their use to tasks where their accuracy has been validated. On the other hand, Rudin (2019) would design interpretable models from the outset. Such models might describe a phenomenon in the same way as people explain things to each other, for example, describing a bird by reference to its head, throat, wings, legs, etc. Rudin would use governance mechanisms to enforce compliance with efforts to develop interpretable models. She cautions about the opportunity for corporations to profit from the intellectual property afforded to a black box, a pertinent concern in light of the collaborative efforts between large corporations and organizations involved with public health and welfare. Focusing on an alternative approach to explaining automated decisions, Wachter et al. (2018) propose counterfactual explanations. Such explanations provide a statement of the decision followed by a statement of how the world would have to be for a different result. For example:

> *"You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan"* (ibid., p. 844).

Wachter and colleagues argue that such explanations are useful for three reasons: they provide reasons for a particular decision; a basis for contesting the decision; and guidance on how to alter future behaviour to receive a preferred outcome. However, such explanations are invariably incomplete, since there may be multiple ways to achieve the desired outcome, only one of which is highlighted, even though other ways may be more achievable for the person affected by the decision. Finally, some researchers do not accept the need for a trade-off between accuracy and explainability. Research on Explainable Artificial Intelligence (XAI) aims to "produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (DARPA 2017). With these goals in mind, Adadi and Berrada (2018) state that XAI refers to "the movement, initiatives, and efforts made in response to AI transparency and trust concerns, more than to a formal technical concept" (p. 52140).

In summary, transparent, accurate explanations of some types of ML model are currently out of reach and there is no consensus amongst researchers about a solution. Claims to accuracy may simply reinforce existing biases, which the models have learned from the data on which they were trained. On the other hand, incomplete explanations are misleading and potentially dangerous, and can hinder efforts to establish accountability. Notwithstanding these concerns, there is optimism that an ML approach may revolutionize decision making in a wide range of domains, allowing professionals to make more effective use of resources.

## Addressing the Accuracy versus Explainability Debate

In recognition of the ongoing debate about a trade-off between accuracy and explainability in some types of AI system, NIHR Greater Manchester Patient Safety Translational Research Centre (NIHR PSTRC) and the Information Commissioner's Office (ICO) in the United Kingdom (UK) commissioned Citizens Juries c.i.c. and the Jefferson Centre to investigate this matter by organizing citizens' juries in two UK cities. Citizens' juries can be thought of as a 'microcosm of the public' (Escobar and Elstub 2017) tasked with deliberating on an issue of national importance and making recommendations to inform policy making. Such discussions are facilitated – in this case by the commissioned organizations listed above – while expert witnesses provide evidence to the participants who in turn question (or cross examine) them.

### *Citizens' Juries Formation, Process and Outputs*

The citizen jurors were drawn from over 450 applicants residing in or around two UK urban centres. Each jury comprised 18 members who, in broad terms, were selected to represent the demographic mix of England (according to the 2011 census) in terms of age, gender, ethnicity, educational attainment and employment status (NIHR PSTRC 2019, jury outputs C4). The juries were held over five days in consecutive weeks in February and March 2019. In each case the jury was asked to address a set of questions about how they weighed AI performance against the need for citizens to be provided with an explanation of automated decisions that affect them. Since jury members were not required to have any prior knowledge of AI systems, they heard presentations from, and asked questions of, expert witnesses. They deliberated in groups to explore the jury questions and produced a report of their findings. At the start and end of the jury they answered a couple of further questions which were used to compare how their views about automated decision making had changed during the process.

The jurors were asked to examine four scenarios in which three different types of automated decision system might be used. The scenarios were stroke diagnosis, matching kidney donors to recipients, shortlisting job applicants, and selecting minor criminal offenders for a rehabilitation programme. The three types of automated decision system which the jurors were asked to consider were described as an expert system (system A), conventional machine learning or random forests (system B) and deep learning (system C), which could provide levels of transparency as indicated in table 1.

A collection of 25 data files and reports of the jury proceedings are available on the NIHR PSTRC website (NIHR PSTRC 2019). These data sources include the questionnaires and responses completed by the jurors together with reasons for their answers; presentation slides and links to videos by expert presenters on both sides of the debate; details of the scenarios considered by the jurors; administrative reports dealing with recruitment and organization of the juries and the brief for the expert witnesses; final reports from both juries; and an overall report from the jury organizers. The author of this paper reviewed all of these files and also had access to one of the expert witnesses with whom to discuss the unfolding of events at the juries.

| | A – Expert System | B – Conventional ML | C – Deep Learning |
|---|---|---|---|
| Accuracy | 75% (experienced human level) | 85% (human expert level) | 95% (beyond human level) |
| Transparency | Full explanation | Partial explanation | No explanation |

**Table 1. Automated Decision Systems**

The final report from the jury organizers (NIHR PSTRC 2019, jury outputs C4) presents some rudimentary statistics reflecting jury members' preferences for the type of automated decision system they would like to see in operation in four different scenarios. These findings are reviewed in this manuscript in light of the nuances and controversies surrounding the accuracy versus explainability debate in AI. Given the goal of this research to assess the potential social consequences of the increasing use of AI decision making systems in a range of domains, the author adopted an inductive qualitative analysis approach which examined the available sources looking for expert discussion of i) the performance of AI systems in terms of the four metrics presented earlier and their implications in different contexts; ii) the possibilities for different types of explanation, even for deep learning systems; and iii) the extent, if any, of a required trade-off between accuracy and explainability in AI and its implications for those affected by it. Particular items of interest included the reasons the juries offered for their decisions, the system options they were asked to consider, and the organization of the juries.

## *Findings from the Juries*

The jurors' preferences by system in each scenario are shown in table 2. In both healthcare scenarios, the jurors' votes showed a clear preference for system C (deep learning), prioritizing accuracy over explanation. While voting results were more mixed in the other two scenarios, the combined votes for systems B and C (ML systems) indicated that the majority of jurors were willing to trade off at least some explanation in favour of accuracy in all cases.

The jurors' votes on the final day present a picture that runs counter to concerns raised elsewhere about the ethical issues associated with automated decisions that cannot be explained (Cath 2018; Mendling et al. 2018). This is particularly so in the two high risk healthcare settings. Key reasons given by the jurors for prioritising AI performance[1] in the two healthcare scenarios and hence preferring system C (deep learning) included:

- It is more accurate – higher diagnostic success rate
- Urgent need to treat the problem more important than explaining it
- Witness said most stroke victims don't ask how the stroke was diagnosed
- Kidney patient just needs to know they've got a match
- Greater accuracy reduces no. of failures feeding back into system – ultimately reducing waiting lists

| | A – Expert System | B – Conventional ML | C – Deep Learning |
|---|---|---|---|
| Stroke Diagnosis | 0 | 5 | 31 |
| Kidney Matching | 1 | 2 | 33 |
| Job Applicant Shortlisting | 7 | 20 | 9 |
| Minor Criminal Rehabilitation | 15 | 13 | 8 |

**Table 2. Jurors' Votes for Automated Decision Systems**

[1] Reports from the citizens' juries use the term AI performance to refer to the accuracy of the decisions made by the three types of AI system. When I refer to the findings in these reports, available on the NIHR PSTRC website (NIHR PSTRC 2019), I use the term performance to enable the reader to trace my sources. Clearly, though, in the context of this research the terms performance and accuracy are synonyms.

In the recruitment scenario, opinion was divided across all three systems. Reasons given for prioritising AI performance focused on the business need to recruit the best candidate for the job, while reasons given for prioritising explanation addressed the business need to understand the criteria which define a successful candidate and also for feedback to be provided to unsuccessful candidates in order to allow them to improve. In the recruitment scenario, system B (conventional ML) had the most support overall amongst jurors, and the highest ranking reasons for this preference were:

- Applicant receives some feedback, also reduces time and resources
- It offers a high level of accuracy and gives some explanation/feedback.

In the criminal justice scenario, opinion was also divided across all three systems. Reasons given for prioritising AI performance viewed the accuracy of the decision as paramount since it might determine an individual's future, while reasons given for prioritising explanation focused on the subjective nature of the data, which it was felt might lead to many spurious decisions, and the need for transparency for the offender, the current victim and any potential future victim. In the criminal justice scenario, system A (expert system) had slightly more support overall than either of the ML systems, and the highest ranking reasons for this preference were:

- Is fully transparent, bias can be clearly identified, allowing human intervention/appeal of decision
- More interaction with officers. Feelings and remorse taken into account
- Full explanation will be given for the end result.

In light of their final positions, it is appropriate to examine the jurors' positions earlier in the process to determine their starting point and identify any changes that occurred during the week. At the end of day 2, jurors were asked to document their reasons for prioritizing AI performance over explanation and vice versa. Key reasons they identified for prioritizing performance over explanation were:

- Efficient use of resources
  a) Developments not constrained by human intelligence
  b) Reduced time to process large amounts of data
  c) Complex time-critical decisions made as quickly as possible
  d) Potential for faster discovery of treatments
- Potential for more accurate decision-making, not influenced by emotions
- We already prioritize effects over explanations in our use of medicines.

In the case of prioritizing explanation over performance, jurors' reasons were:

- Ethical concerns
  a) Safety and well-being more important than performance
  b) Explanations needed to develop trust
  c) Should not introduce bias
- Regulatory processes for AI not in place, so faults cannot be investigated uniformly
- Lack of transparency, so program errors more difficult to address, may lead to unforeseen outcomes

Clearly, the jurors identified context as an important issue. On day 2, before the scenarios were introduced, familiar arguments introduced by the expert presenters were evident in their reasoning, but at this stage they were not applied to specific settings. By day 5, most of these arguments were still present, but the extent to which the jurors thought they applied in a particular case varied considerably. Crucially, the vast majority felt that explanations were unnecessary in the two healthcare settings, if this would compromise any degree of accuracy. Key questions that arise at this stage are: why did the jurors believe that criminal justice data are very subjective, and prone to erroneous decision making, such that systems in use must be capable of explanation, while medical data relating to life-threatening conditions are not so, and can be processed more accurately and without the need for explanation. More widely, how might the jurors have responded if all four metrics and their implications (identified earlier) had been introduced, and the possibilities for different types of explanation had been discussed?

## Discussion

In an effort to address the questions identified above, this research sought a reinterpretation of the findings presented in the citizens' juries reports (NIHR PSTRC 2019). The concept of ambivalence and the mechanisms people adopt to address it informed this reinterpretation (Ashforth et al. 2014; Merton 1976). Preliminary insights from this research suggest that at the end of day 2, the two juries' positions on AI and explainability displayed classic signs of ambivalence, that is, their reasons relating to the need for explanations of AI decision-making showed clear, but opposing, orientations (Merton 1976). However, their perceptions of conflicting orientations were likely relieved to some degree by the facilitators' use of splitting (Ashforth et al. 2014), in which the juries' were asked to list separately their cases for prioritizing explanation over accuracy and vice versa, and a similar ordering was adopted for the presentations of the two expert witnesses advocating these positions. On day 5 the jurors were required to choose their preferred system for each scenario, necessitating a resolution of their conflicting sentiments. Since the metrics of precision, recall and F1 score had not been introduced to them, they were not prompted to consider the high cost of false positives and false negatives in some cases. Therefore, in choosing the most accurate systems for the healthcare scenarios, even though they provided no explanation, they focused on achieving the highest number of correct stroke diagnoses and kidney matches. However, strokes and kidney transplants are life-threatening situations, so the consequences of false negatives and false positives respectively would be likely to result in patient deaths. Being unable to provide explanations to their families in these situations would surely be unacceptable, however infrequently they occurred.

In the recruitment scenario, jurors favoured a compromise approach (ibid.), in which some feedback was available together with a good level of accuracy. The counterfactual explanations discussed earlier could provide such a solution. However, since such explanations are necessarily incomplete, the feedback could be misleading, for example, by highlighting a way of improving that was very difficult for the applicant when another alternative (not revealed) could more easily produce a successful outcome. Again, jurors needed more prompting about the drawbacks of partial explanations. The outcome for the criminal justice scenario is particularly interesting, given the jurors' reluctance to trust in the accuracy of a deep learning approach. Evidently, the fact that the system had to make a judgement about whether someone who had committed a minor criminal offence was likely to commit a major crime within the next six months brought personal considerations sharply into focus. Jurors commented on the subjective nature of the data and the likelihood of spurious decisions, suggesting that they did not trust AI systems to make accurate character assessments.

Two further points are worthy of note. By the end of day 2 the jurors had heard presentations from four of the five expert witnesses on: i) the trade-off between AI performance and explainability; ii) data protection and AI; iii) support for prioritizing AI performance over explainability; and iv) support for prioritizing explainability over AI performance. Clearly, the jury organizers started from the position that providing explanations of AI decision-making necessarily impacts AI performance and so a trade-off between accuracy and explanation is necessary. The assumption that such a trade-off is required is not accepted by all researchers, including one of the expert witnesses, as his slides demonstrate (NIHR PSTRC 2019, jury materials B1, p.37). However, in adopting this approach, the organizers highlighted a potentially negative consequence of providing explanations, thereby bolstering the case for accuracy. Second, the claim made for system C – that some AI systems significantly exceed the accuracy of human experts' decision-making – is a long way from the real-world situation for AI performance, particularly for deployment in critical areas such as healthcare. The legitimacy of the system C option was dealt with in the official juries' report published in May 2019 (more than two months after the juries took place) (NIHR PSTRC 2019, jury outputs C4). However, the jurors' rationales for their system preferences show clearly how telling the accuracy claims were in their decision-making. Arguably, then, results from the juries would look significantly different if system C had been excluded, or significantly challenged on grounds of its claims to performance.

## Conclusion

Citizens' juries are a form of participatory action research which involves consulting members of the public about a matter of public policy. The official report of the juries discussed in this paper describes them as a 'form of "deliberative democracy", based on the idea that individuals from different backgrounds and with no special prior knowledge or expertise can come together and tackle a public policy question' (NIHR PSTRC 2019, jury outputs C4, p. 19). The policy matter addressed by the two UK juries concerned the issue

of providing an explanation of an automated decision to individual(s) affected by it. This is an important policy issue, particularly in light of the GDPR in the EU, and since Brexit, the Data Protection Act 2018 in the UK. The juries were generally supportive of automated decisions even when they could not be explained, especially in the two healthcare scenarios they were asked to consider. This position is not just a challenge to GDPR guidance, but at odds with much research on the use of 'black box' models in high risk areas (e.g. Floridi 2018). The jurors' positions prioritised accuracy over explanation. By introducing some additional metrics to accuracy and highlighting the possibilities for different forms of explanation, this research presents an alternative perspective from the citizen jurors on the consequences of increasing deployment of AI systems in various domains of human life. In revealing ambivalence in the jurors' positions and some mechanisms they adopted to address it, this work highlights the potential for adverse social consequences if debate about the relative merits of accuracy and explainability in AI is presented narrowly.

While citizens' juries have merit as a means of consulting the public about policy matters that affect them, a frequent criticism relates to the extent that they actually inform policy making unless legitimizing a political decision already taken behind closed doors (Bryant and Hall 2017). The question of bias in the information presented to the jurors was addressed in the official juries' report and in the end-of-day questionnaires completed by the jurors. Clearly, there is always scope to introduce bias into citizens' juries, not least because they are usually commissioned by bodies that have a particular interest in the outcome. This is not to suggest resisting engagement by the public in decisions that affect them. Rather, work is needed to design studies so that participants without specific knowledge of AI systems are given sufficient information to make informed decisions. A citizen jury approach could be adopted, but with a knowledgeable presider (such as the judge in a criminal trial) who could moderate proceedings and ensure that the jury was adequately informed. This work may also be progressed in the following ways. Theoretically, there is scope for developing an ethical perspective, for example, in relation to the development of trust, and the way accuracy, explainability and related issues such as social inclusion, safety and accountability influence confidence in and acceptance of AI systems. Finally, given the diverse interests involved in the research field, AI developments are an important area for cross-disciplinary collaboration, where it is just as necessary to break down the barriers to collaboration between researchers as it is to enable members of the public to make informed decisions.

## References

Adadi, A., and Berrada, M. 2018. "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (Xai)," *IEEE Access* (6), pp. 52138-52160.

Ashforth, B., Rogers, K. M., Pratt, M., and Pradies, C. 2014. "Ambivalence in Organizations: A Multilevel Approach," *Organization Science* (25:5), pp. 1453-1478.

Bryant, P., and Hall, J. 2017. "Citizens Jury Literature Review." Retrieved 20 April 2020, from https://sharedfuturecic.org.uk/wp-content/uploads/2018/01/Literature-review-on-Citizen-Juries-25.5.2017.pdf

Cath, C. 2018. "Governing Artificial Intelligence: Ethical, Legal and Technical Opportunities and Challenges," *Philosophical Transactions of the Royal Society A* (376:20180080), pp. 1-8.

DARPA. 2017. "Explainable Artificial Intelligence (Xai)." Retrieved 3 May 2021, from https://www.darpa.mil/program/explainable-artificial-intelligence

De Fauw, J., Ledsam, J. R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C. O., Raine, R., Hughes, J., Sim, D. A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P. T., Suleyman, M., Cornebise, J., Keane, P. A., and Ronneberger, O. 2018. "Clinically Applicable Deep Learning for Diagnosis and Referral in Retinal Disease," *Nature Medicine* (24:9), pp. 1342-1350.

Escobar, O., and Elstub, S. 2017. "Forms of Mini-Publics: An Introduction to Deliberative Innovations in Democratic Practice." *Research and Development Note* Retrieved 29 April 2020, from https://www.newdemocracy.com.au/docs/researchnotes/2017_May/nDF_RN_20170508_FormsOfMiniPublics.pdf

EU. 2016. "Regulation (Eu) 2016/679 of the European Parliament." *Official Journal of the European Union L 119, 59* Retrieved 5 November 2018, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2016:119:TOC

Faraj, S., Pachidi, S., and Sayegh, K. 2018. "Working and Organizing in the Age of the Learning Algorithm," *Information and Organization* (28), pp. 62-70.

Floridi, L. 2018. "Soft Ethics, the Governance of the Digital and the General Data Protection Regulation," *Philosophical Transactions of the Royal Society A* (376:20180081), pp. 1-11.

Ghoneim, S. 2019. "Accuracy, Recall, Precision, F-Score & Specificity, Which to Optimize On?" Retrieved 3 September 2021, from https://towardsdatascience.com/accuracy-recall-precision-f-score-specificity-which-to-optimize-on-867d3f11124

Guardian. 2018. "Artificial Intelligence Tool 'as Good as Experts' at Detecting Eye Problems." Retrieved 6 November 2018, from https://www.theguardian.com/technology/2018/aug/13/new-artificial-intelligence-tool-can-detect-eye-problems-as-well-as-experts

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. 2018. "A Survey of Methods for Explaining Black Box Models," *ACM Computing Surveys* (51:5), p. Article 93.

Lipton, Z. 2018. "The Mythos of Model Interpretability," *ACM queue* (16:3), pp. 1-28.

London, A. J. 2019. "Artificial Intelligence and Black-Box Medical Decisions: Accuracy Versus Explainability," *Hastings Center Report* (49:1), pp. 15-21.

Mendling, J., Decker, G., Hull, R., Reijers, H., and Weber, I. 2018. "How Do Machine Learning, Robotic Process Automation, and Blockchains Affect the Human Factor in Business Process Management?," *Communications of the Association for Information Systems* (43), pp. 1-23.

Merton, R. 1976. *Sociological Ambivalence and Other Essays*. New York: The Free Press.

NIHR PSTRC. 2019. "Citizens' Jury Design Documentation, Items A1-A8, Jury Materials B1-B2, Jury Outputs C1-C15." Retrieved 20 April 2020, from http://www.patientsafety.manchester.ac.uk/research/themes/safety-informatics/citizens-juries/

Rudin, C. 2019. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* (1), pp. 206-215.

Shortliffe, E., and Sepulveda, M. 2018. "Clinical Decision Support in the Era of Artificial Intelligence," *Journal of the American Medical Association* (320:21), pp. 2199-2200.

Wachter, S., Mittelstadt, B., and Russell, C. 2018. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the Gdpr," *Harvard Journal of Law and Technology* (31:2), pp. 841-888.