Stakeholders in Explainable AI

Alun Preece and **Dan Harborne**
Crime and Security Research Institute
Cardiff University, UK
Email: {PreeceAD|HarborneD}@cardiff.ac.uk

Dave Braines and **Richard Tomsett**
IBM Research
Hursley, Hampshire, UK
Email: {dave_braines|rtomsett}@uk.ibm.com

Supriyo Chakraborty
IBM Research
Yorktown Heights, New York, USA
Email: supriyo@us.ibm.com

Abstract

There is general consensus that it is important for artificial intelligence (AI) and machine learning systems to be explainable and/or interpretable. However, there is no general consensus over what is meant by 'explainable' and 'interpretable'. In this paper, we argue that this lack of consensus is due to there being several distinct stakeholder communities. We note that, while the concerns of the individual communities are broadly compatible, they are not identical, which gives rise to different intents and requirements for explainability/interpretability. We use the software engineering distinction between validation and verification, and the epistemological distinctions between knowns/unknowns, to tease apart the concerns of the stakeholder communities and highlight the areas where their foci overlap or diverge. It is not the purpose of the authors of this paper to 'take sides' — we count ourselves as members, to varying degrees, of multiple communities — but rather to help disambiguate what stakeholders mean when they ask 'Why?' of an AI.

Introduction

Explainability in artificial intelligence (AI) is not a new problem, nor was it ever considered a solved problem. The issue first came to prominence during the 'knowledge engineering era' of the late 1970s and early 1980s, when the focus was on building expert systems to emulate human reasoning in specialist high-value domains such as medicine, engineering and geology (Buchanan and Shortliffe 1984). It was soon realised that explanations were necessary for two distinct reasons: system development, particularly testing, and engendering end-user trust (Jackson 1999). Because the systems were based on symbolic knowledge representations, it was relatively straightforward to generate symbolic traces of their execution. However, these traces were often complex and hard for developers to interpret, while also being largely unintelligible to end-users because the reasoning mechanisms of the system were unrecognisable to human subject-matter experts. The latter problem led to approaches aimed at re-engineering knowledge bases to make the elements of the machine reasoning more recognisable, and to make the generated explanations more trustworthy (Swartout, Paris, and Moore 1991). These latter, 'stronger' approaches to explainable AI were arrived at only when the knowledge engineering boom was effectively over, so they gained little traction at the time.

The last decade has seen a number of significant breakthroughs in machine learning via deep neural network approaches that has reinvigorated the AI field (LeCun, Bengio, and Hinton 2015). In this generation of AI development, the issue of explainability has again come into focus, though the term *interpretability* is nowadays more commonly used, indicating an emphasis on humans being able to interpret machine-learned models. As in the 1970s and 1980s, there are differing motives between system developers and users in seeking explanations from an AI system: the former want to verify how the system is working (correctly or otherwise) while the latter want assurance that the outputs of the system can be trusted (Ribeiro, Singh, and Guestrin 2016). Unlike classical expert systems, deep neural network models are not symbolic so there is no prospect of generating intelligible 'reasoning traces' at the level of activation patterns of artificial neurons. Consequently, a distinction has been made between interpretability approaches that emphasise *transparency* and those that are *post-hoc* (Lipton 2016). The former are explanations expressed in terms of the inner workings of a model while the latter are explanations derived 'after the fact' from the workings of the model, such as an explanation in terms of similar 'known' examples from the training data.

However, terminology in relation to explainability in modern AI is far from settled. A recent UK Government report on the state of AI received substantial expert evidence and noted, 'The terminology used by our witnesses

This research was sponsored by the U.S. Army Research Laboratory and the UK Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the UK Ministry of Defence or the UK Government. The U.S. and UK Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

varied widely. Many used the term transparency, while others used interpretability or 'explainability', sometimes interchangeably. For simplicity, we will use 'intelligibility' to refer to the broader issue' (UK House of Lords Select Committee on Artificial Intelligence 2017). Others have used the term *legibility* (Kirsch 2017) while recent thinking once again emphasises 'strong' notions of explainability in causal terms (Pearl and Mackenzie 2018). Terminology is further complicated by concerns over the accountability (Diakopoulos 2016) and fairness (O'Neil 2016) of modern AI systems which, while overlapping the issue of end-user trust, extend into ethical and legal domains. These various perspectives and distinct groups of stakeholders have led to the rapid creation of a large and growing body of research, development, and commentary. Recent work seeks to place the field on a more rigorous scientific and engineering basis by, for example, examining axiomatic approaches to model interpretability (Leino et al. 2018; Sundararajan, Taly, and Yan 2017), exploring more sophisticated methods for revealing the inner workings of deep networks (Olah et al. 2018), and arguing for increased use of theoretical verification techniques (Goodfellow, McDaniel, and Papernot 2018).

In summary, today there is a large community focused on the problem of explainable AI, with some seeking to advance the state of the art, others seeking to assess, critique, or control the technology, and still others seeking to exploit and/or use AI in a wide variety of applications. In our own recent work, we examined explainability and interpretability from the perspective of explanation recipients, of six kinds (Tomsett et al. 2018): system *creators*, system *operators*, *executors* making a decision on the basis of system outputs, *decision subjects* affected by an executor's decision, *data subjects* whose personal data is used to train a system, and system *examiners*, e.g., auditors or ombudsmen. We found this *Interpretable to whom?* framework useful in thinking about what constitutes an acceptable explanation or interpretation for each type of recipient. In this paper, we take a slightly different tack, examining the stakeholder communities around explainable AI, and arguing that there are useful distinctions to be made between stakeholders' motivations, which lead to further refinement of the classical AI distinction between developers and end-users.

Four Stakeholder Communities

Developers: people concerned with building AI applications. Many members of this community are in industry — large corporates and small/medium enterprises — or the public sector, though some are academics or researchers creating systems for a variety of reasons including to assist them with their work. This community uses both terms 'explainability' and 'interpretability'. Their primary motive for seeking explainability/interpretability is quality assurance, i.e., to aid system testing, debugging, and evaluation, and to improve the robustness of their applications. They may use open source libraries created for generating explanations; some well-known and widely-used examples include LIME (Ribeiro, Singh, and Guestrin 2016), deep Taylor decomposition (Montavon et al. 2016), influence functions (Koh and Liang 2017) and Shapley Additive Explanations (Lundberg and Lee 2016). Members of the developer community may have created their own explanation-generating code, motivated by an aim to aid practical system development rather than to advance AI theory. In terms of our *Interpretable to whom?* framework, members of the developer community are system creators.

Theorists: people concerned with understanding and advancing AI theory, particularly around deep neural networks. Members of this community tend to be in academic or industrial research units. Many are also active practitioners, though the theorist community is distinguished from developers by their chief motivation being to advance the state of the art in AI rather than deliver practical applications. Members of the theorist community tend to use the term 'interpretability' rather than 'explainability'. The motive to better understand fundamental properties of deep neural networks has led to some interpretability research being labelled 'artificial neuroscience' (Voosen 2017). A well-known early piece of work identified properties of activation patterns, and also how deep neural networks are vulnerable to adversarial attacks (Szegedy et al. 2014). Recent work in this milieu has looked at feature visualisation to better interpret properties of hidden layers in deep networks (Olah et al. 2018). It has also been suggested that such interpretations may provide new kinds of cognitive assistance to human understanding of complex problem spaces (Carter and Nielsen 2017). Membership of this community of course overlaps with the developer community. For example, in the case of an industry researcher who carries out theoretical work on deep neural network technology (theorist) while also applying the technology to build systems (developer). In our 'Interpretable to whom?' framework, members of the theorist community are considered system creators.

Ethicists: people concerned with fairness, accountability and transparency[1] of AI systems, including policy-makers, commentators, and critics. While this community includes many computer scientists and engineers, it is widely interdisciplinary, including social scientists, lawyers, journalists, economists, and politicians. As well as using 'explainability' and 'interpretability', members of this community use 'intelligibility' and 'legibility' as noted in the introduction. A subset of this community will also be members of the developer and/or theorist communities[2] but their motives in seeking explanations are different: for the ethicist community, explanations need to go beyond technical software quality to provide assurances of fairness, unbiased behaviour, and intelligible transparency for purposes including accountability and auditability — including legal compliance in cases such as the European Union's GDPR legislation (Goodman and Flaxman 2016). Our *Interpretable to whom?* framework considers members of ethicist community to be dispersed across all six roles, though the distinct explanation-seeking

[1] 'Transparency' in the common usage of the term rather than the specific usage by (Lipton 2016) and others.

[2] Indeed, professional bodies including ACM, BCS and IEEE all place significant emphasis on recognising ethical, legal and societal issues in software development.

motive of the ethicist community aligns most closely with system examiners, creators, data subjects and decision subjects.

Users: people who use AI systems. The first three communities comprise the vast majority of people who contribute to the growing literature on AI explainability/interpretability, whereas our fourth generally does not. Members of the user community need explanations to help them decide whether/how to act given the outputs of the system, and/or to help justify those actions. This community includes both 'hands on' end-users but also everyone involved in processes that are impacted by an AI system. Consider an insurance company that uses an AI tool to help decide whether and at what cost to sell policies to clients. The end-users of the tool, the director of the company, and the clients are all members of the user community. Again, members of the user community may also be in other stakeholder communities, sometimes in relation to the same AI system; for example, an academic criminologist who has learned how to apply AI technology to create a predictive analytics tool (developer) to assist them in their research (user), while being aware of societal impacts of their work (ethicist). The *Interpretable to whom?* framework places system operators and decision executors in the user community, along with decision subjects.[3]

Engineering and Epistemological Perspectives

Explanation is closely linked to evaluation of AI systems. As noted in the introduction, early AI explanation efforts aimed to help system developers diagnose incorrect reasoning paths. Modern transparent interpretation methods are akin to such 'traces', while post-hoc explanation techniques can be regarded as 'diagnostic messages'. Moreover, explanations speak to issues of user trust and system impact, to the user and ethicist communities. Colloquially, in software engineering, *verification* is about 'building the system right' whereas *validation* is about 'building the right system'. In terms of explanation, verification is mainly associated with transparent techniques; 'glass box' approaches are essential because it matters greatly how the system is built. Validation is more concerned with what the system does (and does not do) and so post-hoc techniques are often useful here.

In line with this thinking, and at risk of overgeneralising, we assert that the developer and theorist communities tend to focus more on verification: the former because they want a system that is 'built right', and the latter because they are interested understanding how the various kinds of deep neural networks work, and what are their theoretical limits. We suggest that the user and ethicist communities are more focused on validation, being more concerned with what an AI system does than about how it is built. This means that the developer and theorist communities tend to focus on transparency-based explanation techniques, while user and ethicist communities value post-hoc techniques.

[3] Arguably, decision subjects will be aligned with the user or ethicist communities, depending on how 'empowered' they perceive themselves to be in relation to the effects of the system outputs.

From an epistemological perspective, we can consider the familiar framing in terms of knowns and unknowns:

Known knowns: for an AI system based on machine learning, these constitute the set of available training and test data. The ability of the system to deal with known knowns is verified by standard testing approaches (e.g., n-fold cross-validation) and reported in terms of accuracy measures. Within the bounds of the known knowns, transparency-based explanation techniques such as deep Taylor decomposition (Montavon et al. 2016) or feature visualisation (Olah et al. 2018) can be used to 'trace' the relationships between features (in input and hidden layers) and outputs. All four stakeholder communities have a clear interest in understanding the space of known knowns, though we would argue that it tends to be the developer constituency that are most focused on this space: maximising system performance within the space, defining the bounds of the space, and widening those bounds as much as is feasible.

Known unknowns: these constitute the space of queries, predictions, or behaviours that the AI system is intended to perform. The accuracy measures produced in system testing (verification) provide an estimate of the ability of the system to deal well with the space of known unknowns. The value of a system to members of the user community is in terms of this ability (otherwise the system is nothing more than a retrieval tool for known knowns). Feedback processes are needed because system system outputs may prove to be invalid at run-time (e.g., the system recommends an action that turns out to be inappropriate) leading to the generation of additional data for the training (known knowns) space. Members of the theorist community are interested in better understanding how AI systems process known unknowns (Olah et al. 2018; Szegedy et al. 2014), and creating improved architectures for doing so.

Unknown knowns: from the perspective of the AI system, these are things that are outside its scope, but known more widely. Some biases of concern to the ethicist constituency fall into this category: a narrowness or skew in the training data results in a model that is 'blind' to particular prejudices (Diakopoulos 2016; O'Neil 2016). Validation is key to revealing such unknown knowns.

Unknown unknowns: these have recently been highlighted as a key concern in AI system robustness (Dietterich 2017), with a variety of methods being proposed to deal with them, including employing a portfolio of models to mitigate against weaknesses in individual models, and creating AI systems that build causal models of the world (Lake et al. 2017) and/or or are aware of their own uncertainty (Kaplan et al. 2018). Clearly, all four communities have reason to be concerned with unknown unknowns: developers in terms of system robustness, theorists in terms of seeking stronger theories and architectures, ethicists in terms of ethical and legal implications of AI system failings, and users in terms of impacts on themselves and their livelihoods.

In software engineering, formal verification techniques have been used to mathematically define the space of knowns — in terms of a system specification — leaving only the unknown unknowns fully excluded from that space. The theorist community is beginning to think along these

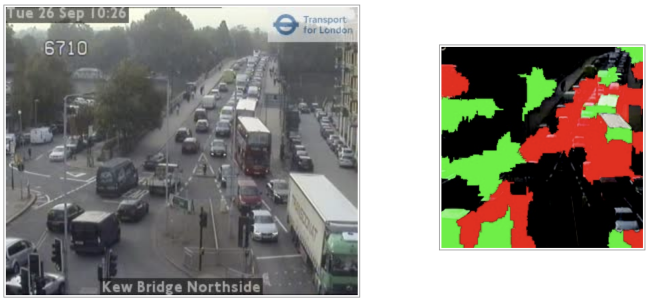

Figure 1: Example saliency map for traffic congestion: the red regions of the input image are most significant in classifying the image as congested

lines (Goodfellow, McDaniel, and Papernot 2018), though how to formally specify the intended behaviour of a deep neural network-based AI system remains an open question. This difficulty has been highlighted in recent years by research into 'adversarial examples', which are designed to fool machine learning models by minimally perturbing input data to cause incorrect classifications (Goodfellow, Shlens, and Szegedy 2014; Szegedy et al. 2014). Such examples take advantage of the difficulty in learning correct classification decision boundaries from limited, high-dimensional data. While several methods to mitigate against such attacks have been proposed (Papernot et al. 2015; Ross and Doshi-Velez 2017), none amounts to a formal verification of the model's behaviour on adversarial inputs (though see (Dvijotham et al. 2018)). Building uncertainty awareness into models so that they can recognise and explicitly deal with such unknown unknowns may be a reliable way of improving system robustness (Gal and Smith 2018), though with unkown implications for human interpretability.

Explanation Types and Discussion

Transparency-based explanations: The definition of transparency in (Lipton 2016) appears consistent with the notion of 'full technical transparency' in (UK House of Lords Select Committee on Artificial Intelligence 2017). Both sources conclude that achieving full transparency is not realistic for anything other than small models, e.g., shallow decision trees or rule bases. A more limited form of transparency is exhibited by attribution techniques that visualise activations in the input or hidden layers of a network (e.g., deep Taylor decomposition (Montavon et al. 2017), feature visualisation (Olah et al. 2018)) often as a *saliency map* showing the features of the input that had most significance in determining the output. While noting that the visualisation element of these approaches is a post-hoc technique (Lipton 2016), we nevertheless consider these methods *transparency-based*, to distinguish them from 'purely post-hoc' approaches that do not derive at all from inner states of the model.

Figure 1 shows an example saliency map for a traffic congestion monitoring system, from (Harborne et al. 2018)[4].

[4]The example map was generated using the LIME software (Ribeiro, Singh, and Guestrin 2016) which does not conform to our definition as being *transparency-based* because it generates a local approximation of the learned model; it is included here only as an example of what a saliency map looks like in general.

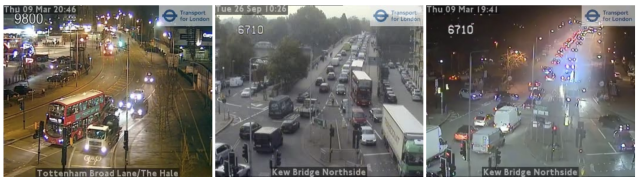

Figure 2: Example explanation-by-example for a traffic congestion classification: the input image is in the middle; the left and right images are training examples with congestion classification probabilities slightly lower and higher, respectively, than the input

From a system verification perspective, such explanations would seem of immediate value to the developer and theorist communities, though with the caveat that many attribution methods are unstable (Sundararajan, Taly, and Yan 2017) and/or unreliable (Kindermans, Hooker, and Adebayo 2017). In addition to these technical concerns, attribution visualisations can be hard to interpret by members of the user and ethicist community where the explanation does not clearly highlight meaningful features of the input. Therefore, such explanations are in danger of making members of these communities *less* inclined to trust the system because they appear to reveal a system that operates in an unintelligible, unstable, 'inscrutable' or 'alien' manner. Even when an explanation seems 'convincing' because it highlights meaningful and plausible features, there is a danger of confirmation bias in the receiver unless counterfactual cases are also included. Providing detailed transparency-based explanations may also overwhelm the recipient — more information is not necessarily better for user performance (Marusich et al. 2018).

Post-hoc explanations: A commonly-used type of post-hoc explanation is approximation using a local model, e.g., visualised as a saliency map as in LIME (Ribeiro, Singh, and Guestrin 2016), or in the form of a decision tree (Craven and Shavlik 1996). Such techniques provide explanations that appear similar to those generated by transparency-based techniques and, if offered to users or ethicists, it is important to communicate clearly that they are actually post-hoc approximations. Explanations in terms of examples — see Figure 2 — are a traditional approach favoured by subject-matter experts (Lipton 2016) and therefore especially appropriate for the user and ethicist communities. Approaches here include identifying instances from the training set most significant to a particular output (Koh and Liang 2017) and employing case-based reasoning techniques to retrieve similar training examples (Caruana et al. 1999). Such approaches have an advantage that counterfactual examples can also be provided. Another common post-hoc technique targeted towards users is to generate text explanations; the approach in (Hendricks et al. 2016) uses background domain knowl-

edge to train the system to generate explanations that emphasise semantically-significant features of the input.

Layered explanations: From the above discussion, it may seem that the sensible approach is to offer different explanations tailored to the different stakeholders, but can we envisage instead a composite *explanation object* that packs together all the information needed to satisfy multiple stakeholders, and can be unpacked (e.g., by accessor methods) per a recipient's particular requirements. Moreover, we can view such an object being layered as follows:
Layer 1 — traceability: transparency-based bindings to internal states of the model so the explanation isn't entirely a post-hoc rationalisation and shows that the system 'did the thing right' [main stakeholders: developers and theorists];
Layer 2 — justification: post-hoc representations (potentially of multiple modalities) linked to layer 1, offering semantic relationships between input and output features to show that the system 'did the right thing' [main stakeholders: developers and users];
Layer 3 — assurance: post-hoc representations (again, potentially of multiple modalities) linked to layer 2, with explicit reference to policy/ontology elements required to give recipients confidence that the system 'does the right thing' (in more global terms than Layer 2) [main stakeholders: users and ethicists].
Example — wildlife monitoring system: Layer 1 (traceability): saliency map visualisation of input layer features for classification 'gorilla'; Layer 2 (justification): 'right for the right reasons' semantic annotation of salient gorilla features; Layer 3 (assurance): counterfactual examples showing that images of humans are not miss-classified as 'gorilla'.

Conclusion

In this paper we have attempted to 'tease apart' some of the issues in explainable AI by focusing on the various stakeholder communities and arguing that their motives and requirements for explainable AI are not the same. We related notions of transparent and post-hoc explanations to software verification and validation, and consideration of knowns/unknowns. We suggested that a 'layered' approach to explanations that incorporates transparency with local and global post-hoc representations may serve the needs of multiple stakeholders.

On a final note, the most influential of our four stakeholder communities is the users — the one that's barely represented in the literature — because, as in the 1980s, failure to satisfy users of AI technology in the long run will be the most likely cause of another 'AI Winter'. Unfulfilled expectations and/or a smaller-than-hoped-for market will lead to investment drying up.

References

Buchanan, B., and Shortliffe, E. 1984. *Rule Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley.

Carter, S., and Nielsen, M. 2017. Using articial intelligence to augment human intelligence. *Distill*. 10.23915/distill.00009.

Caruana, R.; Kangarloo, H.; Dionisio, J.; Sinha, U.; and Johnson, D. 1999. Case-based explanation of non-case-based learning methods. In *Proceedings of the AMIA Symposium*, 212–215.

Craven, M., and Shavlik, J. 1996. Extracting tree-structured representations of trained networks. In *Neural Information Processing Systems (NIPS)*, 24–30.

Diakopoulos, N. 2016. Accountability in algorithmic decision making. *Communications of the ACM* 59(2):56–62.

Dietterich, T. G. 2017. Steps toward robust artificial intelligence. *AI Magazine* 38(3):3–24.

Dvijotham, K.; Stanforth, R.; Gowal, S.; Mann, T.; and Kohli, P. 2018. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'18.

Gal, Y., and Smith, L. 2018. Sufficient conditions for idealised models to have no adversarial examples: a theoretical and empirical study with Bayesian neural networks. *arXiv preprint arXiv:1806.00667*.

Goodfellow, I.; McDaniel, P.; and Papernot, N. 2018. Making machine learning robust against adversarial inputs. *Communications of the ACM* 61(7):56–66.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Goodman, B., and Flaxman, S. 2016. European Union regulations on algorithmic decision-making and a "right to explanation". In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 26–30.

Harborne, D.; Willis, C.; Tomsett, R.; and Preece, A. 2018. Integrating learning and reasoning services for explainable information fusion. In *Proc 1st International Conference on Pattern Recognition and Artificial Intelligence (ICPRAI)*.

Hendricks, L. A.; Akata, Z.; Rohrbach, M.; Donahue, J.; Schiele, B.; and Darrell, T. 2016. Generating visual explanations. In *European Conference on Computer Vision (ECCV 2016)*, 3–19. Springer.

Jackson, P. 1999. *Introduction to Expert Systems*. Addison-Wesley Longman, 3rd edition.

Kaplan, L.; Cerutti, F.; Sensoy, M.; Preece, A.; and Sullivan, P. 2018. Uncertainty aware AI ML: Why and how. *AAAI Fall Symposium Series*.

Kindermans, P.-J.; Hooker, S.; and Adebayo, J. 2017. The (un)reliability of saliency methods. *arXiv preprint arXiv:1711.00867*.

Kirsch, A. 2017. Explain to whom? Putting the user in the center of explainable AI. In *Proceedings of Comprehensibility and Explanation in AI and ML (CEX 2017)*.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Bahavioral and Brain Sciences*.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *Nature* 521(7553):436–444.

Leino, K.; Li, L.; Sen, S.; Datta, A.; and Fredrikson, M. 2018. Influence-directed explanations for deep convolutional networks. *arXiv preprint arXiv:1802.03788*.

Lipton, Z. C. 2016. The mythos of model interpretability. In *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*, 96–100.

Lundberg, S., and Lee, S.-I. 2016. An unexpected unity among methods for interpreting model predictions. In *NIPS Workshop on Interpretable Machine Learning in Complex Systems*.

Marusich, L. R.; Bakdash, J. Z.; Onal, E.; Yu, M. S.; Schaffer, J.; O'Donovan, J.; Höllerer, T.; Buchler, N.; and Gonzalez, C. 2018. Effects of information availability on command-and-control decision making: Performance, trust, and situation awareness. *Human Factors* 58(2):301–321.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2016. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65:211–222.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65:211–222.

Olah, C.; Satyanarayan, A.; Johnson, I.; Carter, S.; Schubert, L.; Ye, K.; and Mordvintsev, A. 2018. The building blocks of interpretability. *Distill*. 10.23915/distill.00010.

O'Neil, C. 2016. *Weapons of Math Destruction*. Crown.

Papernot, N.; McDaniel, P. D.; Wu, X.; Jha, S.; and Swami, A. 2015. Distillation as a defense to adversarial perturbations against deep neural networks. *arXiv preprint arXiv:1511.04508*.

Pearl, J., and Mackenzie, D. 2018. *The Book of Why: The New Science of Cause and Effect*. Allen Lane.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)*, 1135–1144. ACM.

Ross, A. S., and Doshi-Velez, F. 2017. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. *arXiv preprint arXiv:1711.09404*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*.

Swartout, W.; Paris, C.; and Moore, J. 1991. Explanations in knowledge systems: design for explainable expert systems. *IEEE Expert* 6(3):58–64.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tomsett, R.; Braines, D.; Harborne, D.; Preece, A.; and Chakraborty, S. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. In *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*.

UK House of Lords Select Committee on Artificial Intelligence. 2017. AI in the UK: ready, willing and able?

Voosen, P. 2017. How AI detectives are cracking open the black box of deep learning. *Science*.