AISB 2021 Symposium Proceedings:

Overcoming Opacity in Machine Learning

Annual Convention of the Society for the
Study of Artificial Intelligence and Simulation of Behaviour

April 8th, 2021

Edited by Carlos Zednik & Hannes Boelsen

Table of Contents

Preface: Overcoming Opacity in Machine Learning

Carlos Zednik[1] **and Hannes Boelsen**[2]

1 Themes of the Symposium

AI systems developed using machine learning (ML) are notoriously *opaque*: it is difficult to know why they do what they do or how they work [1]. Opacity arises not only because many of these systems are high-dimensional and complex, but also because software developers do not themselves determine the parameter values that drive their behavior [2].

Opacity gives rise to the *Black Box Problem*: an AI system's behavior cannot be readily explained, interpreted, or understood. This problem affects many different stakeholders in the *machine learning ecosystem* [3]: software developers cannot effectively improve the performance of systems whose inner workings they do not fully understand; end-users are unlikely to trust the outputs of a system they are unable to interpret; regulatory bodies are unable to identify the causes of, and thus possibly prevent, unwanted behavioral tendencies such as algorithmic bias. The problem is exacerbated as AI systems become increasingly prevalent in safety-critical domains such as transportation, medicine, policing, finance, and others.

The Black Box Problem has given rise to significant concerns about the proliferation of AI in society. These concerns have spurred ethical debates about the kinds of AI technology that are really worth wanting [4], as well as legal efforts to mandate transparency in algorithmic decision-making [5]. At the same time, these concerns must be balanced against the assumption that machine learning methods may bring considerable gains in efficiency, reliability, accuracy, and overall performance.

The desire to retain the transformative potential of machine learning while mitigating the negative effects of opacity has led to the birth of a new research program: *Explainable Artificial Intelligence* (XAI). Some investigators expect to evade the Black Box Problem by developing ML methods that are "inherently interpretable" [6]. Others, in contrast, intend to solve the Black Box Problem by developing analytic techniques with which to develop "post-hoc explanations" [1]. No matter the approach, the overall aim is to overcome opacity in machine learning.

Like any nascent research program, XAI faces foundational uncertainties: What exactly is opacity, whom does it affect, and how? What is explanation, and how can it be delivered? What, if any, are the limits of explanation in this particular context? These questions must be answered if opacity is to be overcome eventually. For this reason, they constitute the central themes of this symposium.

2 Contributions to the Symposium

The symposium brings together researchers from higher education and industry, spanning the disciplines of artificial intelligence and philosophy. In a total of seven contributions, these researchers introduce several current XAI methods and AI application domains. In so doing, they introduce the technological state-of-the-art and address a focal point of current philosophical investigation.

Four contributions consider the way in which opacity might be overcome in Explainable AI. Whereas Kathleen Creel provides a pragmatist analysis of "explanation" in the XAI context generally, the others focus on post-hoc explanation specifically. Lok Chan investigates the extent to which post-hoc methods can adequately explain a computing system's behavior without faithfully representing its internal structure. David Watson evaluates existing post-hoc methods and identifies a lacking commitment to the quantification of uncertainty. Eunjin Lee, Harrison Taylor, Liam Hiley, and Richard Tomsett introduce two specific problems for post-hoc explanation: a lack of robustness, and a lack of representational fidelity.

Three contributions focus on opacity itself, and thus, on the scope and limits of XAI. Considering the role of machine learning methods in scientific research, Julie Schweer, Paul Grünke, and Rafaela Hillerbrand introduce the notion of "epistemic risk", and evaluate the extent to which models developed using machine learning can, despite their opacity, be used as scientific tools. Andrés Paéz focuses on human interaction with robotic agents, viewing attempts to increase transparency in this context as antithetical to the aim of increasing trust in human-robot interaction. Finally, Vincent Müller distinguishes between different kinds of opacity, and considers their relative impact on ethical debates about data privacy.

These contributions yield an improved understanding of opacity itself, as well as of recent attempts to overcome opacity in Explainable Artificial Intelligence. Although brief, these contributions have the potential to significantly advance current research in artificial intelligence and philosophy alike.

Aknowledgements

The symposium organizers are grateful to the AISB and convention organizers for housing the symposium, especially during the uncertainties of the Covid-19 pandemic. They also thank the members of the program committee for their help with selecting contributed papers: Colin Allen, Cameron Buckner, Nir Fresco, Frank Jäkel, Holger Lyre, Saskia Nagel, and Constantin Rothkopf. Lastly, they are grateful to the numerous contributors and symposium attendees, the Otto-von-Guericke-University and Eindhoven University of Technology for institutional support, and the Deutsche Forschungsgemeinschaft for financial support under grant ZE-1062/4-1.

References

[1] Zednik, C. (2019). Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence. *Philosophy & Technology.*

[2] Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1), 205395171562251.

[3] Tomsett, R., Braines, D., Harborne, D., Preece, A., & Chakraborty, S. (2018). Interpretable to Whom? A Role-based Model for Analyzing Interpretable Machine Learning Systems. ArXiv, 1806.07552.

[4] Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, July-December, 1-21.

[5] Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. International Data Privacy Law, 2017.

[6] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

Function and User-Satisfaction in Explainable AI

Kathleen A. Creel[1]

Abstract. In order for a machine learning model to be useful, it must be used. Opaque models that predict or classify without explaining are often ignored. Thus measuring the satisfaction of those who receive an explanation is one natural way to measure the value of that explanation. A satisfied user will be more likely to trust and therefore to use the system. Nevertheless, I argue that user satisfaction with an explanation alone is not a good metric for the value of that explanation, proposing instead benchmark tests based on accuracy and suitability of explanations for purposes.

1 EXTENDED ABSTRACT

In applied machine learning, success at providing an explanation is often measured by ease and frequency of use among the constituent population. For example, when doctors do not trust decision-assisting software, they will not use it in clinical practice even when they are told that its use decreases diagnostic errors [9]. Since studies have shown that doctors value explanations for the diagnosis very highly, especially step-by-step rationales for decisions, the developers of these systems sought to provide explanations of the kind desired by the doctors [18].

Doshi-Velez and Kim [4] formalized this implicit standard of explainability, equating explanation with satisfaction among the target population. They propose a benchmark measure against which future purported explainers will be measured, similar to the benchmark tests used in Human-Computer Interaction (HCI) [1]. The success of any new proposed explanation or tool for explaining would be measured using human-subject trials. Human subjects would be asked to rate the quality of various short explanations on the same topic. Unbenownst to them, some of the explanations will be written by humans, a pool that will serve as the benchmarks of success. Other explanations will be written by the new automated explainer to be tested. Using such a benchmark, proposed explainers may be compared against each other and against the gold standard of human-generated explanations.

When such a human-subject trial is not possible for financial or other logistical reasons, Doshi-Velez and Kim suggest that those attempting to provide explanations can rely on a default ranking of interpretability (here used synonymously with explanability) of types of models previously established by human-subject trials. For example, they suggest that a sparse linear model or a decision tree can function as a more-interpretable proxy model for an opaque neural net. Thus although token-level tests are best, the general type or class of model can also be ranked in order of its interpretability.

Doshi-Velez and Kim's work has been widely cited [7, 8] and has been followed by similar work using human subject trials to measure the success of various methods for providing interpretability or explanation [13, 15]. I will call these approaches generally the "satisfaction account" of explanation and argue that it has serious problems.

First, classes of models cannot be ranked by their interpretability. A simple decision tree is easy to create and structurally clear when small. However, a thousand-node decision tree is uninterpretable, as it often relies on variables or features that seem to a human to be grue-like or so tiny as to be indistinguishable from one another. Scale matters for the interpretability of a model. There is too much interclass variability for a simple ranking of classes of models to be a good guide to interpretability. A ranking also presupposes a goal-agnostic or model-independent explanation. In order to argue that creating a post-hoc decision tree or fitting a linear model to an opaque algorithm will always improve its explainability, the satisfaction view must posit a unique metric of similarity. But explainability, as with other metrics of comparison, is goal relative [3, 17]. An explanation that suits one purpose may not suit another. Likewise different kinds of opacity may prompt different forms of explanation [2, 19].

The second problem with the satisfaction account is that user preference is the criterion of evaluation. There is no room in the satisfaction account for distinguishing between a feeling of understanding and genuine understanding, let alone the epistemic achievement of a scientific explanation, although doing so is foundational to most accounts of both understanding and explanation [6, 12, 5, 16, 10]. Of course, idealization and abstraction and the fictions they bring with them have long been acknowledged to be an important part of the creation of cognitively tractable explanations [8, 14]. But without access to any information about the functioning of the system, users have no way to determine whether anything about the explanation is factive. Without this knowledge, it is impossible for the user to know whether their trust is warranted.

A third and related concern is that many strategies such as visualization simplify complex models in a misleading way, but may boost their scores with respect to the benchmark merely by relying on human preference for images. For example, visual explanations in machine learning, such as heat maps or saliency methods, often prioritize local explanations over general features and can be fragile to minor and non-adversarial perturbations in the input [11]. However they are rated as being highly explanatory [15]. Likewise, if an image classifier has suggested that an image contains a dog, pointing out the pixels of the original image which most contributed to the diagnosis of "dog" does not answer the why and how of neural network image classification, object detection, and feature localization. On one construal, it explains correct performance on the task by pointing to the algorithms correct performance circular at best. However, ostension may not even uniquely explain correct performance, depending on the counterfactual or contrastive requirements one holds for an explanation. Imagine a classifier suggests with 99% confidence that an image contains a dog. Its second choice, with 78%

[1] Stanford University, USA, email: kcreel@stanford.edu

confidence, might be wolf, relying on many of the same pixels. Such an explanation lacks the implicit contrastive value of many human generated explanations. The lack of a criterion beyond subjective understanding leads to an over-reliance on visualization and condones misleading but prima facie plausible explanations.

For all these reasons, I argue that user-satisfaction human subject trials will not solve the problem of choosing between types of explanation. However, the original goal of the satisfaction theorists should not be abandoned. Benchmark tests and suites have proved to be useful in spurring and measuring progress in other domains. Thus I propose alternate benchmark tests for explainability based on functional explanations, or matching of explanations to goals, in domains where the explanation produced can be compared with insight into the true functioning of the machine learning model or system.

ACKNOWLEDGEMENTS

Thanks to Carlos Zednik for organizing the AISB 2020/2021 Symposium “Overcoming Opacity in Machine Learning.”

REFERENCES

[1] Pedro Antunes, Valeria Herskovic, Sergio F. Ochoa, and Jose A. Pino, ‘Structuring Dimensions for Collaborative Systems Evaluation’, *ACM Comput. Surv.*, **44**(2), (March 2008).

[2] Kathleen A. Creel, ‘Transparency in Complex Computational Systems’, *Philosophy of Science*, **87**(4), (October 2020).

[3] David Danks, ‘Goal-dependence in (Scientific) Ontology’, *Synthese*, **192**(11), 3601–3616, (January 2015).

[4] Finale Doshi-Velez and Been Kim, ‘Towards a Rigorous Science of Interpretable Machine Learning’, *arXiv preprint arXiv:1702.08608*, (2017).

[5] Catherine Z. Elgin, ‘True Enough’, *Philosophical Issues*, **14**(1), 113–131, (October 2004).

[6] Michael Friedman, ‘Explanation and Scientific Understanding’, *The Journal of Philosophy*, **71**(1), 5, (January 1974).

[7] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, ‘Explaining Explanations: An Overview of Interpretability of Machine Learning’, in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, (2018).

[8] Mary B. Hesse, ‘Models In Physics’, *The British Journal for the Philosophy of Science*, **4**(15), 198–214, (November 1953).

[9] Matthew Hutson, ‘Self-taught Artificial Intelligence Beats Doctors at Predicting Heart Attacks’, *Science*, (2017).

[10] Kareem Khalifa, *Understanding, Explanation, and Scientific Knowledge*, Cambridge University Press, October 2017.

[11] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim, ‘The (Un)reliability of Saliency Methods’, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, eds., Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, 267–280, Springer International Publishing, Cham, (2019).

[12] Philip Kitcher, ‘Explanatory Unification’, *Philosophy of Science*, **48**(4), 507–531, (December 1981).

[13] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An Evaluation of the Human-Interpretability of Explanation, 2019.

[14] Ernan McMullin, ‘Galilean Idealization’, *Studies in History and Philosophy of Science Part A*, **16**(3), 247–273, (September 1985).

[15] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation, 2018.

[16] Wesley C. Salmon, *Four Decades of Scientific Explanation*, University of Pittsburgh Press, 2006.

[17] Andrea I Woody, ‘Re-orienting Discussions of Scientific Explanation: a Functional Perspective’, *Studies in History and Philosophy of Science Part A*, **52**, 79–87, (2015).

[18] L. Richard Ye and Paul E. Johnson, ‘The Impact of Explanation Facilities on User Acceptance of Expert Systems Advice’, *MIS Quarterly*, **19**(2), 157, (June 1995).

[19] Carlos Zednik, ‘Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence’, *Philosophy & Technology*, (December 2019).

Beyond Opacity – Epistemic Risks in Machine Learning

Julie Schweer[1], **Paul Grünke**[2] and **Rafaela Hillerbrand** [3]

Extended abstract

Machine learning models play an increasingly important role in a multitude of areas and for a wide range of application contexts. Yet, it is often stated that a central challenge lies in their epistemic opacity or black box nature.

In our contribution, we suggest that besides exploring ways to render machine learning models more transparent, it is fruitful to shed light on sources of epistemic risks that arise in the context of machine learning practices. More precisely, we indicate that machine learning models are epistemically challenging not only because they are often 'opaque' but also because their construction and usage involves dealing with a variety of epistemic risks. Approaching machine learning models from an 'epistemic risk perspective' can, or so we suggest, help addressing them not only as black boxes whose 'inner working' one might aim to reveal but also as tools or artefacts that are intended to serve certain purposes and whose construction (and usage) involves drawing on certain assumptions and making various decisions (cf. e.g. [3]). Seeing how (and at which points) epistemic risks are involved might help to better clarify as to how they can help achieve specific epistemic aims (e.g. make predictions or provide understanding).

Our understanding of epistemic risk differs from the classical notion of inductive risk (as e.g. in [5]). Rather than mainly focusing on the risk of wrongly accepting or rejecting a hypothesis, we follow Justin Biddle and Rebecca Kukla in broadly defining epistemic risk as any risk of error that can arise at various points in knowledge practices ([2], 218). While commonly drawing on the conception of 'epistemic risk' seems particularly motivated by the aim to examine the roles of values in scientific practice, we think that it more generally offers a fruitful framework in order to address challenges arising in the construction and usage of (machine learning) models as the notion of 'epistemic risks' does not (simply) hint to the fact that in various steps epistemic errors may occur, but that even though they may be unavoidable, the epistemic risks can be managed.

Epistemic risks may occur e.g. during problem identification, the operationalization of concepts, data choice, or algorithm design (see [2], p. 220f. or [1]). For example, consider the process of data selection, preparation, and preprocessing in machine learning. Here, several decisions need to be made that involve epistemic risks: What data should be taken as input? How should we deal with redundancies, noise or incomplete data?

Now an interesting question is as to how precisely both of the aforementioned challenges – namely, that (a) machine learning systems are often in some sense opaque and that (b) their construction and usage involves epistemic risks – are related to each other.

For example, think of cases in which the aim is to acquire understanding of real-world phenomena. Here, a common worry is that due to their opacity, complex machine learning models provide only limited understanding of the respective phenomenon under investigation. Yet, Emily Sullivan [6] has recently argued that what hinders us from understanding real-world phenomena by means of machine learning models is not primarily their opacity or lack of transparency but the extent to which the 'link' between model and target phenomenon is uncertain.

Consider, for example, the case of machine learning models used in medical diagnostics. If the aim is to acquire understanding of real-world phenomena such as the development of certain diseases, it seems not only important to learn about the factors relevant for the model to make a particular prediction but also to be certain enough that these factors indeed reflect the actual key-drivers for the development of the respective disease. As Sullivan ([6], 18) puts it, "[w]e want some indication that the model is picking out the real difference makers for identifying a given disease and not proxies, general rules of thumb, or artefacts within a particular dataset."

This suggests that when it comes to an understanding of real-world phenomena, it might be worthwhile to go beyond the functioning of a given model. As an example, Sullivan discusses the case of the so-called 'deep patient model' (see [4]). Taking electronic health records of patients as input, the deep patient model shall help predicting the development of diseases. Yet, despite its remarkable accuracy in predicting certain medical conditions, Sullivan ([6], 18-19) argues that link uncertainty is still present and that therefore, the model provides only limited understanding of why e.g. a particular patient might develop a particular disease.

Moreover, Sullivan points out that the model had difficulties with predicting some diseases such as cases of diabetes mellitus without complications ([6], p. 18f., see also [4], 8f.). The scientists' hypothesis for explaining this was that diabetes mellitus is often diagnosed during general routine checkup tests and that thus, the frequency of these tests does not reliably help predict the occurrence of diabetes ([4], 8-9; see also [6], 18). As Sullivan ([6], 19) puts it, this indicates that the "model in part tracks proxies of disease development, such as previous physicians' decisions to carry out a diagnostic test".

This brings us back to the conception of epistemic risk. Obviously, the decision to rely on a certain kind of data (here: health records)

[1] Institute for Technology Assessment and System Analysis, Karlsruhe Institute of Technology, Germany, email: julie.schweer@kit.edu

[2] Institute for Technology Assessment and System Analysis, Karlsruhe Institute of Technology, Germany, email: paul.gruenke@kit.edu

[3] Institute for Technology Assessment and System Analysis, Karlsruhe Institute of Technology, Germany, email: rafaela.hillerbrand@kit.edu.

comes with certain risks (here, for example: tracking proxies rather than causes of disease development). Drawing on the framework of epistemic risk and examining sources of epistemic risks can, or so we suggest, help address the question of how the model is ‘linked’ to its target. This, in turn, may provide a very first step to approaching the question of how much understanding these models provide.

REFERENCES

[1] J. B. Biddle, ‘On predicting recidivism: Epistemic risk, tradeoffs, and values in machine learning’, *Canadian Journal of Philosophy*, 1–21, (2020). https://doi.org/10.1017/can.2020.27.

[2] J. B. Biddle and R. Kukla, ‘The geography of epistemic risk’, in *Exploring Inductive Risk: Case Studies of Values in Science*, eds., K. Elliott and T. Richards, 215–237, Oxford University Press, (2017).

[3] J. A. Kroll, ‘The fallacy of inscrutability’, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **376**(2133), 20180084, (2018). http://dx.doi.org/10.1098/rsta.2018.0084.

[4] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, ‘Deep patient: an unsupervised representation to predict the future of patients from the electronic health records’, *Scientific reports*, **6**(1), 1–10, (2016). 10.1038/srep26094.

[5] R. Rudner, ‘The scientist qua scientist makes value judgments’, *Philosophy of science*, **20**(1), 1–6, (1953).

[6] E. Sullivan, ‘Understanding from machine learning models’, *British Journal for the Philosophy of Science*, (2019). https://doi.org/10.1093/bjps/axz035.

Explainable AI as Epistemic Representation

Lok Chan

Abstract. As a way to increase transparency for black box models, post hoc but interpretable models have been proposed as "explanations" for these opaque models' behaviors. However, the supposed relationship between these models has been criticized, especially given the structural dissimilarity between them. I propose that theories of representation in philosophy of science can address how these post hoc models can *faithfully represent* another model without completely replicating the structure of the target model. I argue that the philosophical framework of epistemic representation, which concerns itself with how representation mediates our inference of its target, provides a rigorous way in which we can evaluate the success of Explainable AI in the broader context of ethics and transparency.

1 Introduction

Increasing awareness of the potential biases and inaccuracies in artificial intelligence/machine learning has prompted calls for greater transparency in the deployment and development of so-called black box models. A model is considered a black box when its structure is inaccessible or opaque to people, due to the model being either prohibitively complex or a proprietary product. Because of these considerations, the concept of explanation has taken a prominent role in recent research in AI. "Explainable AI" (XAI) refers to a set of methods that, in order to increase transparency and trustworthiness of black box models, construct post hoc but interpretable models that represent black box models in some capacity and serve as the explanation for the behaviors of the original models [4]. Nevertheless, this model of explanation has been criticized on the ground that post hoc models never completely reproduce the structure of the target model [8].

This raises an important philosophical question: must the relationship between the two models be evaluated in terms of similarity? In this paper, I critically examine this problem and propose that philosophical theories of representation can address how these post hoc models can *faithfully represent* another model without completely replicating the structure of the target model. To be clear, my purpose is not to provide a blanket *defense* of XAI; rather, the point is to suggest how the philosophical framework of epistemic representation, which concerns itself with how representation mediates our inference about its target, provides a richer way to evaluate whether or not a particular proxy model has adequately represented its target. The upshot is an account that will allow us to specify with greater rigor the criteria for misrepresentation.

2 Interpretability and Explainability

Let us draw a distinction between models that are interpretable and those that are explainable [8]. Interpretable models are characterized by their native intelligibility. For instance, decision trees are typically considered to be interpretable, as they make predictions based on decision rules that are understandable to humans.

XAI, on the other hand, has been taken to refer to the model's capability of being explained by another post hoc model that is interpretable (in the sense described). Somewhat counterintuitively, explainability is generally a feature of a black box model, i.e., a model whose structure is natively inaccessible to people, as usually only an opaque model would require explaining. Model opacity could occur for two reasons. First, a model can be a black box when its formal structure is beyond human comprehension due to its complexity, or when the structure is simply too disanalogous to human reasoning. Deep learning models, such as convolutional neural networks, are often black boxes due to their deeply recursive structure. Second, a model could become a black box, not because it is uninterpretable, but because the company or corporation that owns the model maintains that its inner workings are a trade secret. For instance, a model called 'Correctional Offender Management Profiling for Alternative Sanctions' (COMPAS), which is widely used in parole decisions in the U.S. justice system to predict an inmate's recidivism risk, is a black box due to its proprietary nature, even though experts have speculated that it is technically an interpretable model [1].

To make a black box model explainable, then, another model has to serve as an "explanation." For example, one proposed explanation of neural networks is to extract a decision tree model, which is considered to be interpretable, by incorporating the outcome of the target neural networks as the training set [9]. Local Interpretable Model-agnostic Explanations (LIME) are proposed to explain complex models with a simplified and interpretable model that is built specifically around one particular prediction and the set of similar cases [7]. Another proposal is to explain black boxes by using counterfactual explanations, which constructs a model that will provide information about how a black box model could have made a different classification or prediction, had the features been different [12]. The essential feature of these XAI methods is that they aim to generate an interpretable model without presupposing knowledge about the internal workings of the black box model [4].

3 The Dissimilarity Argument

One particular relevant view is from the computer scientist Cynthia Rudin, who in a recent paper criticizes the use of post hoc models to explain black box models. Consider the argument in the passage below [8].

> Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would

be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space.

The intuition behind this argument seems to be a conceptual link between fidelity and similarity. A completely faithful post hoc model here seems to mean for any set of input, the post hoc model would give the same output as the target model, so the maximal fidelity here is assumed to be maximal similarity, which, in turn, is assumed to be identity. From this, the idea that "explanations must be wrong" follows as a matter of definition, since any post hoc model that does not make identical prediction is a misrepresentation.

Let us call this the "dissimilarity argument", as the key assumption here is that "to faithfully represent" is taken to mean "to perfectly resemble". Based on this strict equivalence between resemblance and representation, the rest of the conclusion follows: representational fidelity is to be measured in terms of resemblance, and anything short of complete resemblance entails misrepresentation and untrustworthiness.

The difficulty with assuming a strong relationship between representation and similarity is that no model except itself can be a faithful representation. This has two undesirable consequences. First, an argument *against* XAI that presupposes such a strong assumption is somewhat self-defeating, as a definition that renders by fiat all proxy models "wrong" is tantamount to begging the question against the XAI project. Second, this leaves no room for the possibility that there could be trustworthy representation of systems that are very dissimilar to their targets, which, as we shall see in the next section, is a philosophically untenable position.

To be clear, the point is not just a semantic one, but to make explicit the normative implications that the notion of representation has on how these proxy models ought to be evaluated. Both proponents and critics of XAI, I think, would agree that we would benefit from an account of model representation that allows us to specify, with reasonable precision, criteria under which a proxy model can be regarded as a *good* representation or a *mis*representation of its target. Crucially, then, the following questions must be answered.

- Must a representation's fidelity to the target model be measured solely in terms of likeness, resemblance, or similarity?
- Must a representation completely resemble its target in order to be trustworthy?

I now turn to philosophical theories of representation for possible answers.

4 Representation, Fidelity, and Transparency

Philosophical discussion concerning the relationship between representation and resemblance dates as far back as Plato [6]. More recently, in his book *The Language of Art*, Nelson Goodman famously argues against the often assumed view that "A represents B to the extent that A resembles B" [5]. Goodman objects to this definition by pointing out conceptual distinctions between resemblance and representation. Resemblance is *reflexive*: as a matter of fact, the object that most resembles x is x itself. Yet, representation lacks reflexivity, since it seems absurd to assert that someone is a representation of herself. Resemblance is also symmetric: if X resembles Y, then Y resembles X, but representation, unlike resemblance, is *asymmetric*: a painting of a historical figure, say, George Washington, represents him, but Washington does not represent the painting.

The relaxation of the resemblance requirement opens the door to a fruitful inquiry into how representation can help us overcome the opacity of black box models. I suggest that XAI can be seen as a form of *epistemic representation*: representation that facilitates the acquisition of knowledge about the target system through a vehicle of representation [3]. A classic example originates from Rudolf Carnap: a series of connected dots on a subway map represents a subway system by emphasizing the structure as an interconnected network [2]. Even though the subway map does not bear identical or proportional physical properties, such as distance and size, to the actual subway system, it nevertheless allows us to reason about the target system by virtue of being an adequate representation of it. For instance, I can draw valid inferences about the relative locations of the stations based on the map alone. Representation enables what some philosophers call *surrogative reasoning*, which is the act of learning about a target, such as a subway system, by drawing a parallel from a vehicle of representation, such as the subway map [10].

The notion of surrogative reasoning opens the door to a richer framework with which we can evaluate the adequacy of a proxy model as a representation of its target. First, the evaluation of a representation requires us to identify features of the target system the proxy model is intended to highlight [11]. Thus, a good representation must allow us to draw reliable and analogous inferences about these features. This also forges a link between model representation and model transparency. The proxy modeler's intention to use the proxy model to highlight some aspects of the system of interest is also open to critical assessment: she must be able to provide reasons as to why she has chosen to highlight certain features but not others.

REFERENCES

[1] Tim Brennan, William Dieterich, and Beate Ehret, 'Evaluating the predictive validity of the compas risk and needs assessment system', *Crim. Justice Behav.*, **36**(1), 21–40, (January 2009).

[2] Rudolf Carnap, *The Logical Structure of the World: Pseudoproblems in Philosophy*, University of California Press, 1967.

[3] Gabriele Contessa, 'Scientific representation, interpretation, and surrogative reasoning', *Philosophy of science*, **74**(1), 48–68, (Jan 2007).

[4] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, 'Explaining explanations: An overview of interpretability of machine learning', in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, p. 80–89, (Oct 2018).

[5] Nelson Goodman, *Languages of Art: An Approach to a Theory of Symbols*, Hackett Publishing, 1976.

[6] Plato, *Plato: Theaetetus and Sophist*, Cambridge University Press, November 2015.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Why should i trust you? explaining the predictions of any classifier', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144, (2016).

[8] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence*, **1**(5), 206–215, (May 2019).

[9] G. J. Schmitz, C. Aldrich, and F. S. Gouws, 'Ann-dt: an algorithm for extraction of decision trees from artificial neural networks', *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, **10**(6), 1392–1401, (1999).

[10] Chris Swoyer, 'Structural representation and surrogative reasoning', *Synthese*, **87**(3), 449–508, (June 1991).

[11] Bas C van Fraassen, *Scientific Representation: Paradoxes of Perspective*, Oxford University Press, 2008.

[12] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual explanations without opening the black box', *Harvard Journal of Law Technology*, **31**(2), 842–887, (2018).

No Explanation without Inference

David S. Watson[1]

Abstract. Complex algorithms are increasingly used to automate high-stakes decisions in sensitive areas like healthcare and finance. However, the opacity of such models raises problems of intelligibility and trust. Researchers in interpretable machine learning (iML) have proposed a number of solutions, including local linear approximations, rule lists, and counterfactuals. I argue that all three methods share the same fundamental flaw – namely, a disregard for *severe testing*. Techniques for quantifying uncertainty and error are central to scientific explanation, yet iML has largely ignored this methodological imperative. I consider examples that illustrate the dangers of such negligence, with an emphasis on issues of scoping and confounding. Drawing on recent work in philosophy of science, I conclude that there can be no explanation – algorithmic or otherwise – without inference. I propose several ways to severely test existing iML methods and evaluate the resulting trade-offs.

1 Introduction

Machine learning (ML) is increasingly ubiquitous in modern society. Complex algorithms are widely deployed in private industries like finance [3], as well as public services such as healthcare [20]. Their prevalence is driven by results. ML models outperform humans not just at strategy games like chess [17], but at important scientific tasks like antibiotic discovery [19] and tumor diagnosis [10].

High-performance algorithms are often opaque, in the sense that it is difficult for humans to understand the internal logic behind individual predictions. This raises fundamental issues of trust. How can we be sure a model is right when we have no idea why it predicts particular values? While model interpretation is by no means a new concern in statistics, it is only in the last few years that a dedicated subfield has emerged to address the issues surrounding algorithmic opacity.

Interpretable machine learning (iML) comprises a diverse collection of technical approaches intended to render statistical predictions more intelligible to humans [11]. My focus here is on model-agnostic, post-hoc local methods, which explain the individual predictions of some target model without making any assumptions about its form. Prominent examples include local linear approximators (e.g., SHAP [6]), which produce feature attributions that sum to the explanandum; rule lists (e.g., Anchors [15]), which provide explanations via sequences of if-then statements; and counterfactuals (e.g., MACE [4]), which identify one or several nearest neighbors on the opposite side of a decision boundary. Despite their merits, all three approaches fail to meet the severity criteria outlined in Sect. 2. I illustrate the issues with this failure in Sect. 3, and propose some directions for improvement. I conclude in Sect. 4 with a reflection on the trade-offs implied by this analysis.

2 Severe Testing

Mayo [9, 8, 7] argues that the problem of induction is defeasibly resolved by severe testing. The basis for this resolution is her severity principle, which states that "We have evidence for a claim C just to the extent it survives a stringent scrutiny. If C passes a test that was highly capable of finding flaws or discrepancies from C, and yet none or few are found, then the passing result, x, is evidence for C" [7, p. 14]. On Mayo's view, the justification for believing a given hypothesis is a function not of the hypothesis itself or the data it purportedly explains, so much as the tests it has passed. When tests are sufficiently sensitive (i.e., likely to detect true effects) and specific (i.e., likely to reject false effects), then we say they are *severe*.

To make matters concrete, consider a single parameter location test. Let Θ denote the parameter space, and let T be a test that decides between $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_1$, where Θ_0 and Θ_1 are some partition of Θ. We observe sample data x and compute sufficient statistic $d(x)$, which measures the disagreement between x and H_0. Test T rejects H_0 when $d(x)$ meets or exceeds the critical value c_α. We say that H_0 passes an (α, β)-severe test T with data x if:

(S1) $d(x) < c_\alpha$; and
(S2) with probability at least $1 - \beta$, if H_1 were true, then we would observe some sufficient statistic $d(x')$ such that $d(x') \geq c_\alpha$.

Readers well-versed in frequentist inference will recognize some familiar concepts here. The critical value is indexed by the type I error rate α, such that, under H_0, the rejection region of statistics greater than or equal to c_α integrates to α. Under H_1, the rejection region of statistics less than c_α integrates to the type II error rate, β. The complement of this value, $1 - \beta$, denotes the power of the test. A test with small α is said to be specific, since it only accepts hypotheses that are likely to be true; a test with small β is said to be sensitive, since it is able to detect even slight deviations from the null.

While this explication is faithful to the frequentist framework that Mayo favors, the severity criteria are in fact very general, and have been reformulated along Bayesian lines [2]. ML is not inherently aligned with any particular interpretation of probability, and nothing in the proceeding argument depends upon one's preferred method of inference. The epistemological upshot of Mayo's analysis is that science advances knowledge not just by falsifying theories, as Popper would have it [12], but by subjecting hypotheses to increasingly severe tests. Hypotheses earn their warrant by passing such tests, thereby providing positive justification for successful theories.

3 Severity and iML

An algorithmic explanation is an empirical claim relating certain factors in the input data to the resulting prediction. Since empirical claims are typically the realm of science, we may justifiably wonder

[1] University College London, London, United Kingdom. Email: david.watson@ucl.ac.uk

whether Mayo's severity criteria can be fruitfully applied in this setting. I argue that they can and should. I highlight two ways that algorithmic explanations mislead when severity criteria are not taken into account: through ambiguity of scope and sensitivity to confounding.

Local explanations are constructed to apply only in some fixed region of the feature space. Yet iML methods do not generally provide information about the bounds of a given explanation or goodness of fit within the target region. For illustration, I will focus on linear approximators, but the point applies more broadly.

If you zoom in far enough to any point on a continuous function, you will eventually find a linear tangent. This is the intuition behind methods like LIME [14] and SHAP [6]. However, when the regression surface or decision boundary around the target point is extremely nonlinear, the linear region tends to be very small and the estimated coefficients highly unstable. In this case, feature attributions are acutely sensitive to regional bounds. In a simple two-dimensional example, Wachter et al. [21] visually demonstrate how a linear explanation for the same model prediction may assign positive, negative, or zero weight to a feature depending on the scope of the linear window (see Fig. 1).

The most obvious statistical solution here would be to augment iML outputs with information regarding the scope and fit of the approximation. It is common, for instance, in linear regression to compute the significance and standard error of model coefficients. This would satisfy (S1). Power analysis typically requires parametric assumptions or data simulations, which could be used to satisfy (S2). Unfortunately, these strategies are not readily available to algorithms like LIME and SHAP, which use unconventional sampling techniques, kernel weights, and regularization penalties that preclude easy analytic solutions for calculating expected error rates. Nonparametric resampling methods could help but at major computational cost. The problem becomes especially acute as the number of explananda increases.

Another challenge for iML arises when features are highly dependent. The issue can be especially nefarious when auditing for algorithmic bias. If a sensitive attribute is associated with a permissible variable (e.g., if race is well predicted by zip code) then the latter can serve as a proxy for the former. This allows bad actors to get away with discrimination, so long as they can fool an auditor into believing they were using the permissible variable rather than the sensitive one. The concern is not merely speculative. Authors have exploited these vulnerabilities to make discriminatory models pass algorithmic audits [18] and appear fair in user studies [13].

Severe testing cannot, on its own, prevent bad actors from engaging in discriminatory behavior. However, it can make it harder for them to get away with it by elucidating the uncertainty associated with algorithmic explanations under confounding. Just as standard errors for regression coefficients are inflated by collinear predictors, the severity of particular explanations will tend to decrease with strongly correlated features. Reporting the error rates of given outputs will provide much-needed context for users and regulators alike.

Algorithmic fairness is a complex and contested topic. Dozens of statistical fairness criteria have been proposed [1], while impossibility theorems have shown that most popular definitions are mutually incompatible except in trivial cases [5]. No matter which criteria one adopts for a given application, almost all may be expressed in terms of marginal or conditional independences, which means that classical tests can be used for auditing purposes. Severity therefore has a central role to play in holding people and institutions accountable for their algorithmically mediated decisions.

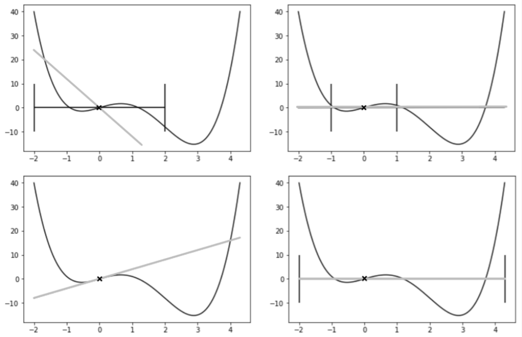

Figure 1: Unstable linear approximations. The grey line in each panel shows a local approximation of the same function centered at the same location. The varying range is indicated by the black bars, leading to vastly different linear explanations. From [21, p. 885].

4 Discussion

Many authors motivate the iML project with appeals to trust. "Why should I trust you?" reads the title of Ribeiro et al.'s paper introducing LIME [14]. "Building trust is essential to increase societal acceptance of algorithmic decision-making," [21, p. 843] write Wachter et al. in their paper on counterfactual explanations. So long as complex algorithms remain opaque, users will harbor suspicions about their reliability in particular cases. That is why we seek transparent explanations that can assuage concerns about unfair or unreasonable model predictions.

But do iML algorithms really settle matters, or merely push the problem one rung up the ladder? After all, why should we trust their outputs? Presumably the target function at least has the advantage of performing well on some test dataset. Can we say the same of algorithms like SHAP, Anchors, or MACE? Their outputs are readily intelligible, and that is clearly a start. But does that necessarily mean that their explanations should all be given equal weight, or are some more reliable than others? How can we be sure that they have not produced unstable estimates or selected the wrong features? Are there principled methods for critically evaluating individual explanations, much like we can critically evaluate individual predictions?

I argue that severe testing holds the key to securing the trustworthiness of algorithmic explanations. The goal of all iML algorithms is to produce claims relating inputs to outputs. Such claims can in principle be tested. That, for instance, is how we come to trust scientific theories – by repeatedly, mercilessly subjecting them to severe tests with quantifiable error rates. There is no good reason to hold iML to a lesser standard.

Concerns over feasibility are legitimate. Bootstrapping methods for evaluating the scope and stability of local explanations could be time consuming. Conditional independence testing, which may aid in fairness audits, is notoriously difficult in high-dimensional settings and provably hard for continuous conditioning events [16]. But if the stakes are sufficiently high that we need an algorithmic explanation in the first place – perhaps even a legally mandated one – then it is important that we get that explanation right.

Proponents of black box algorithms argue that results often matter above all else. Would we prefer a transparent model that diagnoses cancer with 90% accuracy or an opaque one that does so with 99% accuracy? By the same token, we cannot dismiss severe testing for iML merely due to concerns about the computational burden. When consequential decisions depend upon algorithmic explanations, we had better make sure they withstand a stringent scrutiny.

ACKNOWLEDGEMENTS

Thanks to Luciano Floridi and Carl Öhman for their valuable feedback. This research was partially supported by ONR grant N62909-19-1-2096.

REFERENCES

[1] Solon Barocas, Moritz Hardt, and Arvind Narayanan, *Fairness and Machine Learning*, fairmlbook.org, 2019.

[2] Andrew Gelman and Cosma Rohilla Shalizi, 'Philosophy and the practice of Bayesian statistics', *British Journal of Mathematical and Statistical Psychology*, **66**(1), 8–38, (2013).

[3] J B Heaton, N G Polson, and J H Witte, 'Deep learning for finance: deep portfolios', *Applied Stochastic Models in Business and Industry*, **33**(1), 3–12, (2017).

[4] Amir-Hossein Karimi, Gilles Barthe, Borja Balle, and Isabel Valera, 'Model-Agnostic Counterfactual Explanations for Consequential Decisions', in *The 23rd International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 895–905, (2020).

[5] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan, 'Inherent Trade-Offs in the Fair Determination of Risk Scores', in *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67, pp. 43.1–43.23, (2017).

[6] Scott M Lundberg and Su-In Lee, 'A Unified Approach to Interpreting Model Predictions', in *Advances in Neural Information Processing Systems 30*, 4765–4774, (2017).

[7] Deborah Mayo, *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*, Cambridge University Press, New York, 2018.

[8] *Error and Inference*, eds., Deborah Mayo and Aris Spanos, Cambridge University Press, New York, 2010.

[9] Deborah G Mayo, *Error and the Growth of Experimental Knowledge*, University of Chicago Press, Chicago, 1996.

[10] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, and Shravya Ashrafian, Hutan...Shetty, 'International evaluation of an AI system for breast cancer screening', *Nature*, **577**(7788), 89–94, (2020).

[11] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Interpretable*, christophm.github.io/interpretable-ml-book/, Münich, 2021.

[12] Karl Popper, *The Logic of Scientific Discovery*, Routledge, London, 1959.

[13] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton, 'Learning to Deceive with Attention-Based Explanations', in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4782–4793, (2020).

[14] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, '"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, pp. 1135–1144. ACM, (2016).

[15] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin, 'Anchors: High-Precision Model-Agnostic Explanations', in *AAAI*, pp. 1527–1535, (2018).

[16] Rajen Shah and Jonas Peters, 'The Hardness of Conditional Independence Testing and the Generalised Covariance Measure', *Annals of Statistics*, **48**(3), 1514–1538, (2020).

[17] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, and Demis Guez, Arthur...Hassabis, 'A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play', *Science*, **362**(6419), 1140 LP – 1144, (2018).

[18] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju, 'Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods', in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, (2020).

[19] Jonathan M Stokes, Kevin Yang, Kyle Swanson, Wengong Jin, Andres Cubillos-Ruiz, and James J Donghia, Nina M...Collins, 'A Deep Learning Approach to Antibiotic Discovery', *Cell*, **180**(4), 688–702.e13, (2020).

[20] Eric J Topol, *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*, Basic Books, New York, 2019.

[21] Sandra Wachter, Brent Mittelstadt, and Chris Russell, 'Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR', *Harvard Journal of Law and Technology*, **31**(2), 841–887, (2018).

Technical Barriers to the Adoption of Post-hoc Explanation Methods for Black Box AI models

Eunjin Lee[1], Harrison Taylor[2], Liam Hiley[2], Richard Tomsett[1]

Abstract. We examine two key technical barriers to the adoption of post hoc methods for explaining the outputs of black box AI models: their lack of robustness, and the difficulty in assessing explanation fidelity.

1 INTRODUCTION

The recent expansion of Explainable Artificial Intelligence (XAI) research has resulted in a variety of methods that can produce "explanations" for black box AI model outputs. Given the perception that black box methods such as neural networks perform better than more intrinsically interpretable models on some tasks – particularly those involving low level perception (e.g. image recognition, object detection, speech recognition) or natural language processing (e.g. machine translation, text generation) – much XAI work has attempted to create methods that can explain black boxes without having to modify the models themselves. This approach is illustrated in figure 1: the "explanator" is designed to produce a suitable explanation of the black box AI system's output. The explainee uses this explanation to form an interpretation of how the model arrived at that output (by contrast, for an inherently interpretable model, the "explanator" box would not be needed, and the explainee would be able to examine the model directly to interpret how it arrived at a particular output).

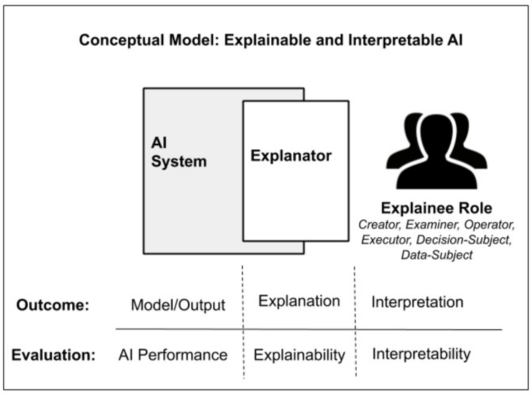

Figure 1. Conceptual framework for XAI. From [1].

In this extended abstract, we ask a key question for such *post hoc* explanation methods [2]: are they good enough? Our discussion expands on the technical issues around XAI identified in [3]. First, we examine some important failings of common explanation techniques that can lead to them producing misleading explanations. Second, we describe recent work evaluating metrics that have been proposed to measure the fidelity of post hoc explanations (i.e. how well the explanations represent the true internal workings of the model being explained). We discuss the implications of these observations in the context of AI systems for supporting high stakes decision making.

2 PROBLEMS WITH ROBUSTNESS

Papers reporting on new explanation techniques rarely evaluate the methods beyond qualitative comparisons with previous methods. Those that perform quantitative evaluations do not use a consistent methodology, so the evaluations are difficult to compare. One result of this is that various flaws in explanation methods have only come to light after extensive testing by other researchers. A particularly popular method that suffers from several serious flaws is LIME: Local Interpretable Model-agnostic Explanations [4]. LIME has been shown not to be robust: given two very similar inputs that result in very similar outputs from the model, LIME is not guaranteed to produce similar explanations [5]. In other work, it has also been shown that LIME could produce very different explanations when run many times on the same input/output pair, due to its reliance on stochastic perturbations of the input data [6]. More generally than just LIME, several explanation methods for neural networks are susceptible to adversarial attacks: it is possible to generate an imperceptibly modified input that produces the same model output, but results in a totally different explanation by the explanation method [7]. Other research has shown the possibility of designing adversarial patches for images that fool both the classifier and the explanation method [8,9] – specifically, the popular Grad-CAM method [10]. These vulnerabilities are perhaps unsurprising given the findings of Adebayo et al. [11], which demonstrated that several explanation methods for neural networks produced explanations that were largely independent of the neural network parameters. This means that they do not do a good job of representing the internal processing that occurs in the network, so cannot be faithful explanations.

3 PROBLEMS WITH ASSESSMENT

While [11] showed that several methods did not produce faithful explanations, it did not attempt to quantify explanation fidelity to allow for quantitative comparisons between methods. Such quantification requires a metric for measuring explanation fidelity. A few such metrics have been proposed; these rely on perturbing input features and measuring the change in output of the model, to see if the explanation has appropriately assigned

[1] Emerging Technology, IBM Research, Hursley, UK. Email: {`ele3`, `rtomsett`}`@uk.ibm.com`.

[2] Crime and Security Research Institute, Cardiff University, Cardiff, UK.

importance to that feature. Recent work tested these metrics on a standard image classifier model [12]. It found that the metrics were statistically unreliable in several aspects: they produced inconsistent rankings of explanation methods between images, such that the average metric score over images is an unreliable indicator of how faithful explanations of future images will be; they produced inconsistent scores between explanation methods, indicating that the metrics were measuring different things despite all purporting to measure fidelity; and they produced different results depending on their implementation details. It is therefore hard to recommend using current metrics as a reliable guide for picking an explanation method, which is particularly problematic as the explanations produced by different methods vary greatly.

4 DISCUSSION

Two important drivers for the pursuit of reliable XAI systems are: (1) to increase trust in AI tools to facilitate a wider adoption of these systems and (2) the increase of regulations which require explanations for accountability purposes. The instability of the post hoc results and the inherent difficulty in quantifiably assessing these methods are concerning for both drivers. Understanding this issue is especially vital in the context of using AI systems which support high stakes decision making.

Consider a scenario where a machine learning system is deployed to assist military personnel in an operation to identify and strike a target [13]. It is crucial that the military personnel have appropriate trust in the system to make an informed decision in the operation. After the event, this decision may be scrutinised at a tribunal and without a reliable XAI infrastructure, it is problematic, both technically and legally, to investigate the decision. This problem is also significant in existing industry scenarios; for example, where AI systems advise medical professionals in diagnosing patients or calculate whether an applicant should be granted a loan. The performance and assessments of current post-hoc methods make them questionable tools to use in these situations.

5 CONCLUSIONS & FUTURE WORK

In this abstract, we have reviewed two key technical issues with explanation methods for black box models: their lack of robustness, and the difficulty in assessing their fidelity. These problems mean that such methods are difficult to trust when used to help decision making, particularly in high stakes scenarios. Indeed, Rudin recommends avoiding black box models entirely in these situations, instead relying on intrinsically interpretable models [3]. Some progress has already been made in creating new neural network models that are *self-explaining* [14, 15], thus making such models inherently interpretable. This is a promising area for future research. However, post hoc explanation methods may still be desirable in situations where a black box model genuinely outperforms the interpretable equivalent, and the use case requires this higher performance. Before developing further explanation techniques, we suggest that the community places more emphasis on appropriately evaluating these methods, ideally with a set of standard benchmarks. This could help drive progress in a similar way to how standardized machine learning benchmarks have contributed to the improvement in machine learning techniques.

ACKNOWLEDGEMENT

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 copy-right notation hereon.

REFERENCES

[1] M. Hall, D. Harborne, R. Tomsett, V. Galetic, S. Quintana-Amate, A. Nottle and A. Preece. A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems. *IJCAI Workshop on eXplainable Artificial Intelligence (XAI)*, (2019).

[2] Z. Lipton. The Mythos of Model Interpretability. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, (2016).

[3] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**:206–215, (2019).

[4] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *Proceedings of the Twenty Fifth ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-16)*, (2016).

[5] D. Alvarez-Melis and T. S. Jaakkola. On the Robustness of Interpretability Methods. *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning (WHI)*, (2018).

[6] E. Lee, D. Braines, M. Stiffler, A. Hudler, and D. Harborne. Developing the sensitivity of LIME for better machine learning explanation. *Proc. SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, (2019).

[7] A. Ghorbani, A. Abid, and J. Zou. Interpretation of Neural Networks is Fragile. *Proceedings of the Thirty Third AAAI Conference on Artificial Intelligence (AAAI-19)*, (2019).

[8] A. Subramanya, V. Pillai, and H. Pirsiavash. Fooling Network Interpretation in Image Classification. *2019 IEEE International Conference on Computer Vision (ICCV)*, (2019).

[9] T. Viering, Z. Wang, M. Loog, and E. Eisemann. How to Manipulate CNNs to Make Them Lie: the GradCAM Case. *Proceedings of the BMVC 2019 Workshop on Interpretable and Explainable Machine Vision*, (2019).

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017).

[11] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems 31 (NeurIPS-18)*, (2018).

[12] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece. Sanity Checks for Saliency Metrics. *Proceedings of the Thirty Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, (2020).

[13] G. White, S. Pierson, B. Rivera, M. Touma, P. Sullivan, D. Braines. DAIS-ITA Scenario. *Proc. SPIE 11006, Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, (2019).

[14] D. Alvarez-Melis, and T. Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. *Advances in Neural Information Processing Systems 31 (NeurIPS-18)*, (2018).

[15] C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin. This Looks Like That: Deep Learning for Interpretable Image Recognition. *Advances in Neural Information Processing Systems 32 (NeurIPS-19)*, (2019).

Robot Mindreading and the Problem of Trust

Andrés Páez[1]

Abstract. Robot mindreading is the attribution of beliefs, desires, and intentions to robots. Assuming that humans engage in robot mindreading, and assuming that attributing intentional states to robots fosters trust towards them, the question is whether the development of mind-readable robots is compatible with the goal of enhancing transparency and understanding in automatic decision making. There is a risk that features that enhance mind-readability will make the mechanisms that determine automatic decisions even more opaque than they already are. And current strategies to eliminate opacity do not enhance mind-readability. This paper discusses different ways of analyzing this apparent trade-off and suggests that a possible solution is to adopt tolerable degrees of opacity that depend on pragmatic factors connected to the level of trust required for the intended uses of the robot.

1 INTRODUCTION

Autonomous Artificial Intelligent Systems (AIS) designed to interact socially with humans are becoming a common presence in our private and public lives. Our increased interaction with personal virtual assistants, social chatbots and, especially, humanoid robots invites the question of how humans interpret, predict and explain their behavior and decisions. The interpretation framework adopted will have practical effects, such as facilitating human-robot interaction and cooperation. But the favored interpretative framework also has implications for the transparency and trustworthiness of AIS, two of the main concerns of software engineers committed to the EPSRC Principles of Robotics [1] and of researchers involved in the explainable AI project (XAI). Whether AIS are interpreted as intentional agents or as purely mechanic devices will affect the perception of transparency and the level of trust placed in them. This paper explores the effects of interpretative frameworks on our trust in AIS and on our ability to understand their decisions. In particular, I want to examine whether transparency and mindreading-based trust are compatible goals in the case of robots. The risk is that by attempting to make AISs more mind-readable, we are abandoning the project of understanding automatic decision processes.

2 WHAT IS ROBOT MINDREADING?

I will use the expression *robot mindreading* to designate the attitude of attributing mental states to AIS in order to explain and predict their decisions and actions. According to the conventional meaning of "mindreading" [2], successful interaction with others involves the attribution of beliefs, desires, emotions, and intentions to make sense of their behavior and predict their future actions. Humans readily attribute mental states to other humans, to non-human animals [3], and even to abstract shapes [4]. It thus seems natural to assume that they also spontaneously attribute mental states to robots.[2] Many researchers have defended the idea that humans naturally engage in robot mindreading [5, 6, 7]. This assumption has been strengthened by recent developments in robot design that aim at facilitating meaningful interaction between robots and humans. The goal for many researchers in robotics is to create multimodal interfaces that closely mimic human appearance, behavior and speech to provide social communicative functionality that is natural and intuitive [8]. Social bots can now evaluate the emotional state of a human and adjust their behavior to build rapport and appear empathic [9]. AIS can also rationalize their decisions by translating their internal state-action representations into natural language [10]. And we must not forget that humans have been primed by pop culture and science fiction to regard robots as autonomous intentional agents. Although there are dissenting views [11, 12], for the purposes of this paper I will assume that humans readily engage in robot mindreading.

3 ROBOT MINDREADING AND TRUST

There is no doubt that the attribution of beliefs, desires and intentions to AIS facilitates human-robot interaction [13, 14]. A gamer's experience will be enhanced if she believes that her artificial opponent has (evil) intentions and desires, and a companion robot will better achieve its purpose if its owner believes that the robot actually cares about his woes. The general idea is that the cognitive and emotional response to robots will be more positive if the user treats it as an intentional agent.

But trust demands more than fluid interaction. Distrust in AIS can take different forms. One source of concern is the widely publicized danger posed by biased algorithms. Governments and the private sector have taken strides to address the ethical challenges posed by AI because they are aware that public trust is essential for the consolidation of the so-called 4th Industrial Revolution. A different source of concern is the perception that decisions made by an automatic system are not reliable, even when unbiased, and should not be trusted. Patients are reluctant to use health care provided by medical artificial intelligence even when it outperforms human doctors [15] and most people do not trust automated vehicles [16].

In human-human interaction, honesty, competence and value similarity are essential to establish both *cognitive* and *emotional* trust [17]. Cognitive trust is based on good rational reasons [18], on one's acquaintance with the trustee, and on evidence about his or her reliability; emotional trust is based on the positive feelings generated by our interactions with others. Honesty, competence

[1] Department of Philosophy and Center for Research and Formation in AI (CinfonIA), Universidad de los Andes, Bogotá, Colombia. Email: apaez@uniandes.edu.co

[2] In the literature on human-robot interaction it has become increasingly popular to talk about taking "the intentional stance" towards AIS. I prefer the more theoretically neutral term mindreading.

and value similarity can only be ascribed to others by attributing to them the adequate intentions and beliefs from which these traits can be inferred. The question is whether people will also be more trustful towards AIS if their decisions and behavior are seen as the result of mental states from which honesty, competence and value similarity can be inferred. Intuitively the answer should be affirmative. If a robot behaves in ways that resemble to a high degree those of a trustworthy human, and if the user makes sense of the AIS's behavior by attributing mental states to it, there is no reason to believe that the user will not trust the AIS. Prima facie, then, enhancing traits that convey intentions and beliefs conducive to the creation of trust should be a goal of robotics.

The main problem with this answer is that it extrapolates the trust-building features of human relations to the field of robotics without having enough empirical support. A meta-analysis of factors affecting trust in human-robot interaction revealed that "robot characteristics, and in particular, performance-based factors, are the largest current influence on perceived trust in HRI" [19, p. 523]. This finding is in line with the performance-based definitions of trust found in the literature on multi-agent systems [20].

The meta-analysis also found that factors related to human attitudes towards robots had a small role in trust building. The authors do not conclude that human factors have no influence on HRI. "Rather, the small number of studies found in this area suggests a strong need for future experimental efforts on human-related, as well as environment-related, factors" [19, p. 523]. It could be argued that this meta-analysis focused only on cognitive trust, ignoring the fact that emotional trust is more likely to be detected as an effect of robot mindreading. There are in fact several studies about the emotional reaction of humans towards robots [21], and there is anecdotal evidence of emotional attachments to robots [22]. However, none of these studies have measured emotional trust as an independent variable, so it is impossible to draw any conclusions about the relationship between mindreading and emotional trust.

In sum, the empirical evidence for the trust-building effects of people's attitudes towards robots, and in particular, of the interpretative framework adopted towards them, is inconclusive. I should add that in most cases humans are plainly aware that they are interacting with an artificial being that lacks intentions, consciousness, desires and free will. Despite attributing mental states to machines as an expedient means to predict and explain their behavior in certain contexts, humans are still able to identify true intentional systems. More importantly, momentary rapport and fluid interaction do not entail overall trust and understanding. Trust is not directed towards the individual decisions of an AIS but rather towards its global performance and towards the object itself. The sense of understanding and trust that arises from attributing mental states to AIS can quickly disappear when the machine behaves in unexpected ways.

4 ROBOT MINDREADING AND OPACITY

For the purposes of this paper, I will assume that robot mindreading builds trust, i.e., that there are certain design features of robots that make it easier for people to attribute to them trust-conducive mental states. Working under this assumption we can now ask if the field of robotics should work towards enhancing mindreading. The main reason for raising this question is that the problem of trust in AI systems has a flip side. According to many recent papers that advance the research agenda of XAI [23, 24, 25, 26, 27, 28, 29], to trust an AIS is to understand its actual decision-making process, to make it explainable, transparent, comprehensible and interpretable.[3] Transparency and trust go hand in hand. The question I want to address is whether this second source of trust is theoretically and practically compatible with the goal of promoting the attribution of trust-conducive mental states to robots. If they are not, which one should prevail? Is it possible to develop them in complimentary fashion?

In many cases, the question of trust in AIS is not accompanied by a demand for transparency. For example, when the goal of a humanoid robot is to provide emotional support, users need to feel that their social companion is empathic and understanding. Otherwise they will stop using it. This is a form of interaction that requires trust, in particular, trust in the judgments, perceptions and advice of the AIS, but it is unlikely that users will feel the need to know how these are reached. In fact, the AIS's utility may be negatively affected by increased transparency [31]. However, since many social robots are used in healthcare environments, providers and regulators will want to know whether the content that the AIS is transmitting to a patient, a child or a senior in a vulnerable emotional or physical condition promotes their emotional wellbeing and is not detrimental to their mental health. Thus, healthcare professionals will seek transparency in the decision-making process of social robots.

Furthermore, according to the 4th Principle of Robotics crafted by EPSRC and AHRC,

> Robots are manufactured artefacts. They should not be designed in a deceptive way to exploit vulnerable users; instead their machine nature should be transparent. … although it is permissible and even sometimes desirable for a robot to sometimes give the impression of real intelligence, anyone who owns or interacts with a robot should be able to find out what it really is and perhaps what it was really manufactured to do [1, p. 127].

Is making robots more mindreadable a violation of this principle? Is mindreadability a kind of deception? The principle allows for robots that give the impression of real intelligence, but at the same time there must be a way to make their decision processes transparent. Is it possible to have it both ways?

Although they both aim at trust-building, transparency and mind-readability are goals that pull in different directions. The purpose of the addition of features that promote the attribution of beliefs and intentions to an AIS is to facilitate the kind of interaction and closeness that leads to emotional trust, and that allows the user to make sense of its decisions. But the search for an explanation for the decisions of an AIS aims at a different goal: to make sure that its decisions are *warranted*. Transparency generates *cognitive trust* in AIS [32, 33]. Chances are that in order to achieve one goal, developers will sacrifice the possibility of achieving the other.

Robot systems are still in their infancy in terms of their ability to accurately explain their own behavior, especially when

[3] Each of these terms has been fleshed out in different ways in the literature. I will use "transparency" as a catch-all term for all of these variants. See [30] for a comprehensive analysis.

confronted with noisy sensory inputs and executing complex sequential decision processes [28]. Attempts to explain a robot's decisions and behavior using data-driven approaches are likely to fail given the noisy inputs [34], but the possibility of designing a "transparent robot" is an ongoing research project with some promising results (see below).

However, the current trend in HRI is to design robots that offer natural language explanations that do not purport to represent their inner state or describe their sequential decision processes. Instead, the idea is to offer the explanation that a human would offer when performing a similar action. This idea has been labeled "explainable agency" [35].

Consider two recent examples of this approach. Ehsan et al. introduce what they call "AI rationalization":

> AI rationalization is a process of producing an explanation for agent behavior as if a human had performed the behavior. AI rationalization is based on the observation that there are times when humans may not have full conscious access to reasons for their behavior and consequently may not give explanations that literally reveal how a decision was made. In these situations, it is more likely that humans create plausible explanations on the spot when pressed [10, p. 81].

There is no intention to make AI rationalizations an accurate representation of the true decision-making process. Instead, rationalization sacrifices accuracy for real-time responses, is more intuitive to non-expert humans and will generate higher degrees of satisfaction, confidence, rapport, and willingness to use autonomous systems.

Hellström and Bensch defend a similar approach in which "understanding a robot" means having a successful interaction with it. And achieving a natural, efficient and safe interaction requires mindreading:

> Understanding of a robot is not limited to physical actions and intentions, but also includes entities such as desires, knowledge and beliefs, emotions, perceptions, capabilities, and limitations of the robot. … Hence, we say that a human understands a robot if she has sufficient knowledge of the robot's [state-of-mind] in order to successfully interact with it. [36, pp. 115-116].

A common assumption of both accounts is that robot mindreading is useful to fulfill the intended purpose of the AIS. But usefulness is an interest-relative notion. Robot mindreading is not useful at all for a developer trying to debug or improve the reliability of a robot. Thus usefulness—or utility—is one of the keys to understanding the relation between transparency and mindreading. For some users it is useful to tolerate a high degree of opacity; for some, it is not useful at all.

Risk is the other key to the relation between transparency and mindreading. If robots are perceived as intentional agents, their actions have real effects on the psyche of their users, as we saw in the case of social robots used in healthcare environments. This means that robot designers have a responsibility towards vulnerable users of the robot that goes beyond the intended goal of providing companionship and entertainment. Their responsibility is to guarantee to a reasonable degree that the actions of the robot will not be detrimental to the patients, and this can only be achieved by understanding the underlying decision processes.

Thus, both dimensions have to be considered in robot design and implementation. A robot can fulfill the utility dimension to a very high degree while also obtaining a high grade on a risk scale. What should the recommended course of action be? There is no algorithm that can help us decide which dimension should prevail. It depends on the kind of utility, the kind of risk, the needs of the users and the risk aversion of the people responsible for the implementation of the robot. Therefore, the resolution of the tension between mindreading and transparency is pragmatic through and through.

Seen from another angle, the relation between mindreading and transparency has an ethical side. The rationalizations offered by a robot are, strictly speaking, a false account of its decision process, but they are offered to the user without disclaimer to make her interaction with the robot easier. In a sense, we have created lying robots. Should this disqualify them as morally worthy companions? Zerilli et al. have expressed their concern "that automated decision-making is being held to an unrealistically high standard here, possibly owing to an unrealistically high estimate of the degree of transparency attainable from human decision-makers" [37, p. 661]. Should we then tolerate the same level of insincerity that we find in human-human interactions?

The philosophy of testimony offers a possible answer to this question. According to the anti-reductionist position about testimony, human communication would be impossible if we did not have a natural tendency to believe what other people say without demanding justification at every junction. In Tyler Burge's words, "a person is a priori entitled to accept a proposition that is presented as true and that is intelligible to him, unless there are stronger reasons not to do so" [38, p. 469]. Among the reasons to doubt a testimony are clear signs of insincerity or incompetence or both. A user interacting with a social robot could also claim a presumptive right [39] to believe the reasons it offers to explain its behavior, unless the reasons are perceived as obviously false or nonsensical.

But this answer is insufficient. In high-stakes situations, such as those encountered in law, finance or medicine, a user will demand that the reasons offered match the underlying decision processes. It will not be enough that the explanations offered make sense and seem true. In these areas, procedure, evidence, statutes, and precedent are necessary elements of a satisfactory explanation. In the parlance of philosophers of testimony, the testimony has to be "reduced" or justified. Robots also have to make high-stakes decisions that require complex explanations not likely to be delivered in the form of friendly chatter or ready-made natural language explanations. The use of real-time graphical outputs to represent the internal states and decision-making processes taking place within a robot seems to be a promising road to robot transparency [28. 40]. This approach does not require the use of mindreading-friendly features. Quite the contrary. By making explicit the robot's software hierarchical architecture, it makes it difficult to think of the robot as a being with human-like mental states.

5 CONCLUSIONS

Many intuitions about our interaction with robots might turn out to be right, but it is important to verify them empirically. My first goal in this paper has been to call attention to the lack of empirical evidence for the success of robot mindreading as a trust-building mechanism. Even if robot mindreading turns out to be an effective way to generate emotional trust, this goal has to be balanced against other competing goals such as transparency and cognitive trust. There is no formula that can determine how to weigh these factors, and it is necessary to acknowledge that pragmatic factors will inevitably decide the way forward.

I do not want to claim that transparency and emotional trust are incompatible in principle. Some authors remain confident that it is possible to create transparent robots that are nevertheless emotionally engaging and useful tools across a wide range of domains [31]. But it is important to recognize that mindreading and transparency are in tension and that there are practical, theoretical, and philosophical obstacles that must be overcome before this tension can be resolved.

REFERENCES

[1] Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., Newman, P., Parry, V., Pegman, G., Rodden, T., Sorrell, T., Wallis, M., Whitby, B., & Winfield, A. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, *29*(2), 124-129.

[2] Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

[3] Mameli, M., & Bortolotti, L. (2006). Animal rights, animal minds, and human mindreading. *Journal of Medical Ethics*, *32*, 84-89.

[4] Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology, 57,* 243–259.

[5] de Graaf, M., & Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series Technical Reports* (pp. 19-26). Palo Alto: AAAI Press.

[6] Dennett, D. (1987). *The intentional stance*. Cambridge: MIT Press.

[7] Pérez-Osorio, J., & Wykowska, A. (2019). Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, DOI: 10.1080/09515089.2019.1688778.

[8] Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, *42*, 177-190.

[9] Novikova, J., & Watts, L. (2015). Towards artificial emotions to assist social coordination in HRI. *International Journal of Social Robotics*, *7*, 77.

[10] Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). Rationalization: A neural machine translation approach to generating natural language explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 81-87). New York: ACM.

[11] Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in human neuroscience*, *6*, 103.

[12] Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. PLoS ONE. https://doi.org/10.1371/ journal.pone.0002597.

[13] Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2002). *A survey of socially interactive robots: Concepts, design, and applications*. Technical Report No. CMU-RI-TR-02-29, Robotics Institute, Carnegie Mellon University.

[14] Fink, J. (2012). Anthropomorphism and human likeness in the design of robots and human-robot interaction. In *International Conference on Social Robotics* (pp. 199-208). Berlin: Springer.

[15] Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650.

[16] Hutson, M. (2017). A matter of trust. *Science*, *358*, 1375-1377.

[17] Gambetta, D. (1988). *Trust: Making and breaking cooperative relations*. Oxford: Basil Blackwell.

[18] Lewis, J., & Weigert, A. (1985). Trust as a social reality. *Social Forces, 63*(4), 967-985.

[19] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human factors*, *53*(5), 517-527.

[20] Witkowski, M., Artikis, A., & Pitt, J. (2001). Experiments in building experiential trust in a society of objective-trust based agents. In *Trust in Cyber-societies* (pp. 111-132). Berlin: Springer.

[21] Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An experimental study on emotional reactions towards a robot. *International Journal of Social Robotics*, *5*(1), 17-34.

[22] Klamer, T., Ben Allouch, S., & Heylen, D. (2011). Adventures of Harvey – Use, acceptance of and relationship building with a social robot in a domestic environment. In: M.H. Lamers, & F. J. Verbeek (Eds), *Human-robot personal relationships*. Berlin: Springer.

[23] Ribeiro, M. T., Singh, S., & 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* (pp. 1135-1144). New York: ACM.

[24] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv: 1702.08608*.

[25] Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv:1708.08296*.

[26] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

[27] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, *51*(5), Article 93.

[28] Edmonds, M., Gao, F., Liu, H., Xie, X., Qi, S., Rothrock, B., Zhu, Y., Wu, Y.N., Lu, H., & Zhu, S. C. (2019). A tale of two explanations: Enhancing human trust by explaining robot behavior. *Science Robotics*, *4*(37).

[29] Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines*, *29*(3), 441-459.

[30] Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, *16*(3), 31-57.

[31] Wortham, R. H., & Theodorou, A. (2017). Robot transparency, trust and utility. *Connection Science*, *29*(3), 242-248.

[32] Witkowski, M., & Pitt, J. (2000). Objective trust-based agents: Trust and trustworthiness in a multi-agent trading society. *Proceedings of the Fourth International Conference on MultiAgent Systems* (pp. 463-464). Boston: IEEE.

[33] Tong X., Zhang W., Long Y., & Huang H. (2013). Subjectivity and objectivity of trust. In: *International Workshop on Agents and Data Mining Interaction. ADMI 2012* (pp. 105-114). Berlin: Springer.

[34] Anjomshoae, S., Najjar, A., Calvaresi, D., & Främling, K. (2019). Explainable agents and robots: Results from a systematic literature review. In: N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (Eds.), *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. Montreal: International Foundation for Autonomous Agents and Multi- Agent Systems.

[35] Langley, P., Meadows, B., Sridharan, M., & Choi, D. (2017). Explainable agency for intelligent autonomous systems. In *Twenty-Ninth IAAI Conference*.

[36] Hellström, T., & Bensch, S. (2018). Understandable robots. What, why, and how. *Paladyn, Journal of Behavioral Robotics*, *9*, 110-123.

[37] Zerilli, J., Knott, A., Maclaurin, J., & Gavagahn, C. (2019). Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32, 661-683.

[38] Burge, T. (1993). Content Preservation. *Philosophical Review, 102*, 457-488.

[39] Fricker, E. (1995). Telling and trusting: Reductionism and anti-reductionism in the epistemology of testimony. *Mind*, *104*, 393-411.

[40] Wortham, R. H., Theodorou, A., & Bryson, J. J. (2017). Robot transpar-ency: Improving understanding of intelligent behaviour for designers and users. In In Y. Gao, S. Fallah, Y. Jin, & C. Lakakou (Eds.), *Towards Autonomous Robotic Systems: 18th Annual Conference, TAROS 2017, Proceedings* (pp. 274-289). Berlin: Springer.

Deep Opacity Undermines Data Protection and Explainable Artificial Intelligence

Vincent C. Müller[1]

Abstract: It is known that big data analytics and AI pose another threat to privacy, and it is known that there is some kind of 'black box problem' in AI. I propose that (1) the 'black box' becomes the 'black box problem' in a context of justification for judgments and actions, crucially in the context of privacy. (2) This will suggest distinguishing two kinds of classic opacity and introducing a third: The subjects may not know what the system does ('shallow opacity'), the analysts may not know what the system does ('standard black box opacity'), or even the analysts cannot possibly know what the system might do ('deep opacity'). (3) If the agents, data subjects as well as analytics experts, operate under opacity, then they cannot provide some of the justifications for judgments that are necessary to protect privacy – e.g. they cannot give "informed consent" or assert "anonymity". It follows from (2) and (3) that agents in big data analytics and AI, often cannot make the judgments needed to protect privacy. So big data analytics makes the privacy problems worse, and the remedies less effective. Closing, I provide a brief outlook on technical ways to handle this situation.

Keywords: big data analytics, black box problem, deep opacity, explainable AI, justification, opacity, privacy

1. Background: Opacity and data protection

1.1. Opacity and justification

It has not been sufficiently explained in the modest literature there is why 'opacity' or 'the black box problem in AI' is an issue, and for whom. I will show that at bottom the phenomenon of epistemic 'opacity' in computational modelling [1], machine learning [2] and data science [3], stems from our human practice of explaining and justifying our actions, especially our judgments. And that is a problem for data protection.

It turns out that the opacity issue is not limited to AI, it appears in data analytics, also; particularly in big data analytics.

1.2. Standard regulations

In the EU, many of these issues have been taken into account with the *General Data Protection Regulation* (GDPR) [4]. The GDPR [4] was agreed in the European Parliament and the EU Council in April 2016 and became law from May 2018 in the member states. Member states have the right to specify further rules that do not contradict this regulation (§10, cf. §13). It is a very powerful regulation that embeds a number of new features, while being the result of extended negotiations. I expect it to be a setting the political debate for a numbers of years and would be surprised if substantial changes of the political and regulatory situation were to occur in the next 10 years. Note that the major prior EU document, 95/46/EC, was approved 21 years earlier, only at the level of a 'directive' for national law [5], whereas the GDPR automatically became law itself. The GDPR foresees a "right to explanation" – the extent to which this goes and to which it can be enforced is disputed, however [6-8]. In any case, an inability to explain decisions appears to violate due process, especially when such decisions are challenged.

I will thus take this regulation as a general indicator on what kind of criteria are needed to act lawfully and, to some extent, ethically in privacy matters. It is not my aim to evaluate this regulation, but rather to take it as "what we've got" and see what kinds of problems we can expect in our special case: big data.

The GDPR can be summarised in the following points:

1. It concerns "Personal Data": Name, address, localisation, online identifier, health information, income, cultural profile, …
2. Communication: Who gets the data, why, for how long? (No use for other 'incompatible' purposes. Use as long as necessary.)
3. Consent: Get clear informed consent
4. Access: Provide access to my data
5. Right to be forgotten (not for research)
6. Right to explanation for contracts (& right to have a person decide)
7. Marketing: Right to opt out
8. Legal: Maintain EU legislation when transferring data out
9. Need for a "data protection officer" in your organisation
10. Impact assessment prior to high-risk processing (new technology, personal information, surveillance, sensitive)

The crucial points for our discussion of opacity are no. 3 (informed consent) and 6 (right to explanation), to a lesser extent 4 (access to my data) and 5 (right to be forgotten). To stress this again, these demands of an exemplary data protection regulation like the GDPR just reflect the demands that we make on human agents: to be able to justify their decisions and actions to some extent.

2. Types of opacity

[1] Philosophy & Ethics Group, TU Eindhoven, Netherlands.
Email: v.c.muller@tue.nl. http://www.sophia.de.

2.1. Shallow opacity: AI as an instrument of power

AI is used in automated decisions systems and decision support systems, especially through 'predictive analytics'. The output of such a system may be relatively trivial like "this restaurant matches your preferences", "the patient in this X-ray has completed bone growth", or have greater significance for a person, e.g. if it says "application to credit card declined", "donor organ will be given to another patient" or "target identified and engaged". At the same time, it will often be impossible for the affected person to know how the system came to this output, i.e. the system is 'opaque' to that person, who is typically also the/a data subject. The institution using the system (e.g. the bank) may just tell me that "the system has decided" but not how and why – even if the institution can find out what the reasons were. In such cases opacity is just a matter of decision from some party that is in power and which could find a justification, if it wanted. The opacity in these cases I shall call 'shallow opacity'. This kind of asymmetric opacity to *users only* is the classic (and important) 'knowledge is power'. The users have no control over output, and thus not responsible for the output. This kind of opacity is not specific to AI, it can happen in any use of data science for a decision or decision-support system.

This has been used to classify the notion of opacity in general: "They are opaque in the sense that if one is a recipient of the output of the algorithm (the classification decision), rarely does one have any concrete sense of how or why a particular classification has been arrived at from inputs." [2], but as we shall see it is really a special case. It appears that shallow opacity is the notion Surden and Williams had in mind when they talk about 'technological opacity': "'technological opacity' applies any time a technological system engages in behaviors that, while appropriate, may be hard to understand or predict, from the perspective of human users." [9].

It is often said that such matters raise "significant concerns about lack of due process, accountability, community engagement, and auditing" [10]. These algorithmic systems are part of a power structure, which is why Danaher talks about an 'algocracy' and concludes that "we are creating decision-making processes that constrain and limit opportunities for human participation" [11]. Again, however, the issue of power structure also applies when the opacity does not concern only the user or data subject – this kind of opacity is the subject of our next section.

2.2. Standard or 'black box' opacity

Many AI systems rely on machine learning techniques in (simulated) neural networks that will extract patterns from a given dataset through 'learning'. These networks are organised in 'layers', one of which is the 'input layer' and one the 'output layer, with one or many 'hidden layers' in between. If there is more than one such hidden layer, the network is often called a 'deep' neural network (DNN), and the learning is 'deep learning' [12]. Data connections either flow between layers in one direction, in 'feed-forward' systems, or in any direction, in 'recurrent' systems. The network is can be recalibrated through a feedback system, which changes the outcome on a given income, i.e. the system 'learns'.

These networks learn broadly in three different ways: supervised, semi-supervised (e.g. reinforcement) or unsupervised – though these ways are not mutually exclusive. In the 'supervised' case the system is told about an output whether it is correct or incorrect, or is shown 'good' outputs ('AlphaGo' was of this sort, which beat a top-ranked Go player in 2016, using a supercomputer with 1920 processors and 280 GPUs [13]). In the 'reinforcement' case it is told about a broader target (e.g. win the game of Go) but not whether an output (e.g. a move) was correct or incorrect ('AlphaGo Zero' is of this sort [14]); finally, in the unsupervised case the system tries to find patterns by itself, without feedback on which patterns are useful or correct (these systems are of central importance in statistics [15]). We may find patterns we were not looking for – that nobody knew, and that we cannot explain.

With these techniques, the 'learning' captures patterns in the data and these are labelled in a way that appears useful to the programmer while the programmer does not really know how these patterns came about: "We can build these models, but we don't know how they work." [16]. In fact the programs are typically evolving, so when new data comes in, or new feedback is given, the patterns in the learning system change. There is a significant recent literature about the limitations of machine learning systems [17, 18], that are essentially sophisticated data filters – and quite possibly at the peak of the 'hype cycle' at the moment. Furthermore, the quality of the program depends heavily on the quality of the data provided, following the old slogan "garbage in, garbage out". So, if the data already involved a bias (e.g. police data about the skin colour of suspects, or job data including gender), then the program will reproduce that bias. There are proposals for a standard description of datasets in a 'datasheet' that would make the identification of such bias more feasible [19].

What this means for our purposes is that the outcome cannot really be explained, it is opaque to the user or programmers, especially but not uniquely in the less supervised learning ways. It this thus more opaque than the cases of 'shallow opacity' above, in that the opacity does not just apply to the users or data subjects; it also applies to the experts – the agent-relativity was stressed by [20]. Opacity for experts is a remark not only about what the experts know at a particular point in time, but also about what they can know, even after research. For that reason, the resulting AI is often called 'black box AI' – it features a black box between input and output rather like the human mind is a black box in the eyes of a (methodological) behaviourist.

This kind of opacity is what is often mentioned in general discussions of opacity, so I call it 'standard opacity'. It is, however, specific to AI, in fact to a particular method of AI, namely machine learning. It features the problems of distribution of power mentioned under shallow opacity, and it shows the inability to provide justification for its output. It is probably what authors have in mind who say that the systems "… are opaque: it is difficult to know why they do what they do or how they work", how to explain their "explanatory success" [20]. The systems know things, but we do not, i.e. in the terminology of the 'Rumsfeld cases' there are "unknown knowns" [21].

Perhaps the issue of democratic legitimacy is more urgent in the case of standard opacity, since it cannot be easily relieved. Kissinger pointed out that there is a fundamental problem for democratic decision-making if we rely on a system that is supposedly superior to mere humans, but cannot explain its decisions. He says we may have "generated a potentially dominating technology in search of a guiding philosophy" [22]. In a similar vein, [23] stresses that we need a broader societal move towards more ‘democratic’ decision-making to avoid AI being a force that leads to a Kafka-style impenetrable suppression system in public administration and elsewhere.

2.3. Deep opacity

What has not been sufficiently taken into account in the discussions of opacity and the ‘black box’ is *to whom* the system is opaque, and *to what extent*. It can be opaque to a user, but not to the programmer. It may be opaque to both, but in a way that the expert analyst can overcome. I suggest there is a third case, where opacity cannot possibly be removed, even for the human expert, even if best efforts are made in the generation of the algorithms: ‘deep opacity’.

In order to introduce deep opacity it will be useful to remind ourselves of the kinds of questions we are supposed to answer, where opacity gets in the way – this follows from the section “Standard regulations” above. We had said “The crucial points for our discussion of opacity are no. 3 (informed consent) and 6 (right to explanation), to a lesser extent 4 (access to my data) and 5 (right to be forgotten).” Some questions that follow from these demands are:

Does this data include information about me?
Can you give me access to all the data about me?
Does this data include personal information?
Does this data include a particular piece of information ‘that p’?
Is the data in this dataset anonymous? Can it be de-anonymised?
What information can be derived from this data?

Data analytics and AI for data are ways to find information that is in the data. Prior to carrying out the analysis, and prior to combining this data with other data, we cannot even know what kind of result the analysis will reveal. It might even reveal patterns that are unknown to the subject of the data itself, if the data is about a person or a company. It will reveal statistical patterns that only allow predictions with a certain degree of certainty – and correlation is not causation. It can be shown formally, that certain sets of queries to a database will reveal the entire database [24] or that the data can be re-identified. No matter how well blended the data soup is, there are ways to de-blend some of it and find some information. Anonymisation is a case in point: A dataset that provably cannot reveal any information about a particular item in it would have to be devoid of information – if there is information, who knows how useful it can be? (E.g. the nationality of individuals may reveal very little, but if a particular one is a rare property, perhaps even unique, it can unravel the whole information about that person.) There is a host of examples with databases that were deemed safe to be released into the public domain – and then a smart way to de-anonymise them was found [25].

The data contains information, but how much? – We cannot know. We can only know what a particular method can reveal on a particular dataset, not what future methods might reveal, or what a dataset might reveal when combined with another dataset. So even under ideal conditions and with perfect expert knowledge, we cannot justify answers to the questions above. That is deep opacity.

3. Ethical judgment under opacity

If an agent operates under opacity (on a set of data) they cannot judge certain things, e.g. whether that data constitutes a threat to privacy. They can also not “give informed consent” to the use of data since being “informed” would require knowing at least the main implications of giving that consent. I cannot be responsible for what I cannot know (which is more than what I can know, which is more than what I do know).

There is a long tradition in ethics that recognises what one knows as an important factor in how one’s actions are evaluated. This tradition is especially important and developed in criminal law, so I will use this standard used there to explain – note that the terminologies used differ in different jurisdictions.

Opacity in big data analytics is truly a dilemma: Big data analytics always has deeper opacity (it is definitional for this field), and if we have deeper opacity, then we do not have standard data protection. So with big data we get the worse of both worlds: Increasing potential to do harm, with less ability to find out what is ethically right and to enforce what we think is right.

4. Conclusion

We have proposed an analysis of opacity in AI and data science that is shaped by the societal context in which the issue arises, namely data protection and justification in automated decision systems. It appeared that there are three main types of opacity, ‘shallow’ opacity which is a matter of power structures; ‘standard’ opacity, which is a matter of processing method; finally ‘deep’ opacity, which is a matter of informational content and robustness in the face of new methods or additional data. It would appear that avoiding the standard and deep varieties of opacity would require massive changes in the practices of data science, in particular the avoidance of identifiable data altogether, as well as openness about data sources, processes and stakeholders – this is not impossible, but it may prove a large demand for this very fruitful new science.

References

[1] P. W. Humphreys, "The philosophical novelty of computer simulation methods," *Synthese*, vol. 196, no. 3, pp. 615–626, 2009.

[2] J. Burrell, "How the machine ‘thinks’: Understanding opacity in machine learning algorithms," *Big Data & Society*, vol. 3, no. 1, pp. 1-12, 2016-06-01 00:00:00 2016, doi: 10.1177/2053951715622512.

[3] J. Symons and R. Alvarado, "Can we trust Big Data? Applying philosophy of science to software," *Big Data & Society*, vol. 3, no. 2, 2016, doi: 10.1177/2053951716664747.

[4] GDPR, "General Data Protection Regulation: Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC," *Official Journal of the European Union*, vol. 119, no. 04.05.2016, pp. 1–88. [Online]. Available: http://data.europa.eu/eli/reg/2016/679/oj

[5] EU Parliament, "Data Protection Directive: Directive on the protection of individuals with regard to the processing of personal data and on the free movement of such data (95/46/EC)," *Official Journal of the European Union*, vol. L281, no. 23.11.1995, pp.

31-50. [Online]. Available: http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046

[6] S. Wachter, B. D. Mittelstadt, and L. Floridi, "Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation," *International Data Privacy Law*. [Online]. Available: http://dx.doi.org/10.2139/ssrn.2903469

[7] B. Goodman and S. Flaxman, "European Union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine,* vol. 38, no. 3, 2017. [Online]. Available: https://arxiv.org/abs/1606.08813.

[8] S. Wachter, B. D. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," *Harvard Journal of Law & Technology,* vol. 31, no. 2, 2018. [Online]. Available: http://dx.doi.org/10.2139/ssrn.3063289.

[9] H. Surden and M.-A. Williams, "Technological opacity, predictability, and self-driving cars," *Cardozo Law Review,* vol. 38, no. 121, pp. 121-181, 2016. [Online]. Available: http://scholar.law.colorado.edu/articleshttp://scholar.law.colorado.edu/articles/24.

[10] M. Whittaker *et al.* "AI Now Report 2018." New York University. https://ainowinstitute.org/AI_Now_2018_Report.html (accessed.

[11] J. Danaher, "The Threat of Algocracy: Reality, Resistance and Accommodation," *Philosophy & Technology,* journal article vol. 29, no. 3, pp. 245-268, 2016, doi: 10.1007/s13347-015-0211-1.

[12] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, Mass.: MIT Press, 2016.

[13] The Economist, "Showdown: Win or lose, a computer program’s contest against a professional Go player is another milestone in AI," *The Economist,* vol. 16.03.2016, 2016. [Online]. Available: https://www.economist.com/science-and-technology/2016/03/12/showdown.

[14] D. Silver *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play," *Science,* vol. 362, no. 6419, pp. 1140-1144, 2018, doi: 10.1126/science.aar6404.

[15] J. Hinton and T. Sejnowski, *Unsupervised learning: Foundations of neural computation*. Cambridge, Mass.: MIT Press, 1999.

[16] W. Knight, "The dark secret at the heart of AI," *MIT Technology Review,* vol. May/June, no. 11.04.2017, 2017. [Online]. Available: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.

[17] G. Marcus. "Deep learning: A critical appraisal." https://arxiv.org/abs/1801.00631 (accessed.

[18] D. Castelvecchi, "Can we open the black box of AI?," *Nature,* vol. 538, no. 7623, pp. 20-23, 2016/10/01 2016, doi: 10.1038/538020a.

[19] T. Gebru *et al.* "Datasheets for datasets." https://arxiv.org/abs/1803.09010 (accessed.

[20] C. Zednik, "Solving the black box problem: A normative framework for explainable artificial intelligence," *Philosophy & Technology,* forthcoming. [Online]. Available: https://arxiv.org/abs/1903.04361.

[21] P. J. Nickel, "The Ethics of Uncertainty for Data Subjects," in *The Ethics of Medical Data Donation*, J. Krutzinna and L. Floridi Eds. Cham: Springer International Publishing, 2019, pp. 55-74.

[22] H. A. Kissinger, "How the enlightenment ends: Philosophically, intellectually—in every way—human society is unprepared for the rise of artificial intelligence," *The Atlantic,* vol. June. [Online]. Available: https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/

[23] S. Cave, "To save us from a Kafkaesque future, we must democratise AI," *The Guardian,* vol. 04.01.2019. [Online]. Available: https://www.theguardian.com/commentisfree/2019/jan/04/future-democratise-ai-artificial-intelligence-power

[24] J. M. Abowd, "How will statistical agencies operate when all data are private?," *Journal of Privacy and Confidentiality,* vol. 7, no. 3, pp. 1-15, 2017.

[25] L. Rocher, J. M. Hendrickx, and Y.-A. de Montjoye, "Estimating the success of re-identifications in incomplete datasets using generative models," *Nature Communications,* vol. 10, no. 1, p. 3069, 2019/07/23 2019, doi: 10.1038/s41467-019-10933-3.