Interpretability and Transparency in Artificial Intelligence

Brent Mittelstadt

Abstract

Artificial Intelligence (AI) systems are frequently thought of as opaque, meaning their performance or logic is thought to be inaccessible or incomprehensible to human observers. Models can consist of millions of features connected in a complex web of dependent behaviours. Conveying this internal state and dependencies in a humanly comprehensible way is extremely challenging. Explaining the functionality and behaviour of AI systems in a meaningful and useful way to people designing, operating, regulating, or affected by their outputs is a complex technical, philosophical, and ethical project. Despite this complexity, principles citing 'transparency' or 'interpretability' are commonly found in ethical and regulatory frameworks addressing the technology. This chapter provides an overview of these concepts and methods design to explain how AI works. After reviewing key concepts and terminology, two sets of methods are examined: (1) interpretability methods designed to explain and approximate AI functionality and behaviour; and (2) transparency frameworks meant to help assess and provide information about the development, governance, and potential impact of training datasets, models, and specific applications. These methods are analysed in the context of prior work on explanations in the philosophy of science. The chapter closes by introducing a framework of criteria to evaluate the quality and utility of methods in explainable AI (XAI), and to clarify the open challenges facing the field.

Keywords

Artificial Intelligence; Machine Learning; Algorithm; Transparency; Interpretability; Explainable AI; XAI; Accountability; Philosophy of Science; Ethics

Introduction

Artificial intelligence (AI) challenges our notions of accountability in both familiar and new ways. Systems we are increasingly entrusting with life-changing decisions and recommendations (e.g. employment, parole, and creditworthiness) have their foundation in our technological past but are now digital, distributed, and often imperceptible. When important decisions are taken which affect the livelihood and well-being of people, one expects that their rationale or reasons can be understood.

Compared to human and organizational decision-making, AI poses a unique challenge in this regard. The internal state of a trained machine learning model can consist of millions of features connected in a complex web of dependent behaviours. Conveying this internal state and dependencies in a humanly comprehensible way is extremely challenging (Burrell 2016; Lipton 2016). How AI systems make decisions may thus be too complex for human beings to thoroughly understand their full decision-making criteria or rationale. Despite the difficulty of explaining the 'black box' of AI, transparency remains one of the most common principles cited in AI ethics frameworks produced internationally by public–private partnerships, AI companies, civil society organizations, and governments (Jobin et al. 2019).

Given these constraints and the importance attached to understanding how AI works (Mittelstadt et al. 2019), it is reasonable to ask how the functionality and behaviour of AI systems can be explained in a meaningful and useful way. This chapter aims to answer precisely this question. The chapter proceeds in six parts. Key terminology, concepts, and motivations behind AI interpretability and transparency are first discussed. Next, two sets of methods are examined: interpretability methods designed to explain and approximate AI functionality and behaviour and transparency frameworks meant to help assess and provide information about the development, governance, and potential impact of training datasets, models, and specific applications. The chapter then turns to prior work on explanations in the

philosophy of science and what this work can reveal about how to evaluate the utility and quality of different approaches to interpretability and transparency. Finally, the chapter closes with a discussion of open challenges currently facing AI interpretability and transparency.

Background

To survey interpretability and transparency in AI, it is essential to distinguish, as far as possible, a set of closely related and overlapping terms. Broadly agreed definitions and boundaries for terms such as 'interpretability', 'transparency', 'explanation', and 'explainability' do not yet exist in the field. As a result, the conventions adopted here may not be universal and may contradict other work.

Nonetheless, we can begin to unpack the topic of explaining AI by examining the different types of questions we may ask about AI systems to make them understandable.

How does an AI system or model function? How was a specific output produced by an AI system? These are questions of *interpretability.* Questions of interpretability address the internal *functionality* or external *behaviour* of an AI system (see below for further explanation of this distinction). A fully interpretable model is one which is human comprehensible, meaning that a human can understand the full set of causes of a given output (Lisboa 2013; Miller 2019). Poorly interpretable models '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' (Burrell 2016: 1). Interpretability can also be defined in terms of the predictability of the model; a model is interpretable if a well-informed person could consistently predict its outputs and behaviours (Kim et al. 2016). Questions of model behaviour narrowly address how a particular output or behaviour of the model occurred.[1] However, model behaviour can also be broadly interpreted to include effects on reliant institutions and users and their AI-

influenced decisions; for example, how a physician's diagnosis was influenced by an expert system's recommendation, are also relevant (High Level Expert Group on Artificial Intelligence 2019: 18).

How was an AI system designed and tested? How is it governed? These are questions of *transparency*. Unlike interpretability, transparency does not address the functionality or behaviour of the AI system itself but rather the processes involved in its design, development, testing, deployment, and regulation. Transparency principally requires information about the institutions and people that create and use AI systems as well as the regulatory and governance structures that control both the institutions and systems. Here, interpretability play a supplementary but supportive role. Interpretable models or explanations of specific decisions taken by a system may, for example, be needed for regulators to effectively audit AI and ensure that regulatory requirements are being met in each context of use.

What information is required to investigate the behaviour of AI systems? This is a question of *traceability*. To audit the behaviour of AI systems, certain evidence is needed, which can include 'data sets and the processes that yield the AI system's decision, including those of data gathering and data labelling as well as the algorithms used' (High Level Expert Group on Artificial Intelligence 2019: 18). This data needs to be consistently recorded as the system operates for effective governance to be feasible. Traceability is thus a fundamental requirement for post hoc auditing and explanations of model behaviour; without the right data, explanations cannot be computed after a model has produced a decision or other output (Mittelstadt et al. 2016). Traceability is, however, outside of the scope of this chapter, which is limited to surveying the landscape of methods for AI interpretability and transparency.

The value of interpretability and transparency

As these questions indicate, interpretability and transparency can be valued for many reasons in AI. Interpretability is not a universal necessity in AI. In low-risk scenarios in which errors have little to no impact or in which predictive performance is the sole concern, knowing how a model *functions* or a particular decision was reached may be irrelevant to the problem being solved. However, in many cases, it may be insufficient to merely receive a reliable prediction; rather, understanding how the prediction was made may also be necessary to reliably solve the problem at hand (Doshi-Velez and Kim 2017; Molnar 2020).

In philosophy of science, 'understanding' is treated as an intrinsic good of explanation (Lipton 2001). Understanding how a model *functions* can be inherently valuable for the sake of scientific discovery, human curiosity, and meaning-making (Molnar 2020). These intrinsic goods can be distinguished from the instrumental value of interpretability and transparency in AI as they support goods such as: (a) implementing accountability and auditing mechanisms; (b) complying with relevant legislation and enabling users to exercise legal rights (Doshi-Velez and Kim 2017; Wachter et al. 2017); (c) debugging and refining models (Kulesza et al. 2015); (d) detecting bias and dangerous behaviours; (e) assessing the societal impact of AI (Mittelstadt et al. 2016; Wachter et al. 2020); (f) encouraging user trust (Citron and Pasquale 2014; Ribeiro et al. 2016; Zarsky 2013); and (g) supporting human workers and institutions to work more effectively with AI systems (Rudin 2019a; Samek et al. 2017). Of course, these instrumental goods need to be balanced in practice against the alleged risks of opening systems to public scrutiny, including risks to intellectual property and commercial secrets, potential gaming of decision-making systems, and exploitation of user trust with deceptive or false explanations (Burrell 2016; Mittelstadt et al. 2019; Wachter et al. 2018).

In a recent report, the UK Information Commissioner's Office and Alan Turing Institute distinguish between six categories of explanations of AI systems (Information

Commissioner's Office, The Alan Turing Institute 2020: 20) according to what is being explained. Specifically, explanations can address (a) the *rationale* for a decision; (b) the *responsibility* for the system's development, management, usage, and user redress; and (c) what and how *data* has been used to reach a decision; as well as steps taken to consider (d) *fairness*; (e) *safety and performance*; and (f) the social *impact* of the decision-making process. This taxonomy speaks to common interests underlying requests for explanations of AI models and decisions. Data explanations, for example, can provide details on the data used to train a model, including its source, collection method, assessments of its quality and gaps, and methods used to clean and standardize the data. Impact explanations can provide individuals with information regarding the potential impact of a system on their interests and opportunities, which can inform their decision to 'use' the system or provide a starting point for investigating the impact of the system across relevant populations (Mittelstadt 2016; Sandvig et al. 2014). These types of explanations can likewise be highly valuable for researchers working with machine learning to evaluate the epistemological validity, robustness, and limitations of models and systems (Franzke et al. 2020: 36–46; Mittelstadt et al. 2019). Standardized forms of disclosure (see the section 'Evaluating the quality of explanations') can provide consistency across such explanations.

There are thus many motivators for making AI more understandable. As these different goods suggest, interpretability and transparency can serve many different stakeholders and interests. Explanations can be offered to expert developers, professionals working in tandem with a system (Berendt and Preibusch 2017), and to individuals or groups affected by a system's outputs (Mantelero 2016; Mittelstadt 2017; Wachter and Mittelstadt 2019). Understanding the different potential goods and risks of interpretability, as well as the needs and interests of relevant stakeholders, is essential to ensure a good match between the methods chosen and local contextual requirements (see section).

Interpretability

Several concepts are common across the questions and goods that motivate interpretability in AI. *Interpretability methods* seek to explain the *functionality* or *behaviour* of the 'black box' machine learning models that are a key component of AI decision-making systems. *Functionality* and *behaviour* are both elements of interpretability. The distinction is effectively one between model processing and its outputs; *functionality* refers to the internal calculations or analysis performed by or within the model, whereas *behaviour* refers to its outputs, which are visible to users and affected parties. Viewing outputs does not strictly require comprehension of the method that produced them, although the latter could certainly help develop a richer understanding of the significance and meaning of the outputs.

Trained machine learning models are *'black boxes'* when they are not comprehensible to human observers because their internals and rationale are unknown or inaccessible to the observer or known but uninterpretable due to their complexity (Guidotti et al. 2018; Information Commissioner's Office, The Alan Turing Institute 2020). *Interpretability* in the narrow sense used here refers to the capacity to understand the functionality and meaning of a given phenomenon, in this case a trained machine learning model and its outputs, and to explain it in human understandable terms (Doshi-Velez and Kim 2017). We will return to broader accounts of interpretability and explanation as philosophical concepts later in the chapter (see section 'Philosophy of explanations').

'Explanation' is likewise a key concept in interpretability. Generically, explanations in AI relate 'the feature values of an instance to its model prediction in a humanly understandable way' (Molnar 2020: 31). This rough definition hides significant nuance. The term captures a multitude of ways of exchanging information about a phenomenon, in this case the *functionality* of a model or the rationale and criteria for a decision, to different stakeholders (Lipton 2016; Miller 2019). Unfortunately, in the literature surveyed in this

chapter, it is often deployed in a conceptually ambiguous manner. Terminological confusion is thus common in the field (Mittelstadt et al. 2019).

To understand how 'explanation' is used in the field of interpretable AI, two key distinctions are relevant. First, methods can be distinguished in terms of what it is they seek to explain. *Explanations of model functionality* address the general logic the model follows in producing outputs from input data. *Explanations of model behaviour*, in contrast, seek to explain how or why a particular behaviour exhibited by the model occurred, for example, how or why a particular output was produced from a particular input. Explanations of model *functionality* aim to explain what is going on inside the model, whereas explanations of model *behaviour* aim to explain what led to a specific behaviour or output by referencing essential attributes or influencers on that behaviour. It is not strictly necessary to understand the full set of relationships, dependencies, and weights of features within the model to explain model behaviour.

Second, interpretability methods can be distinguished in how they conceptualize 'explanation'. Many methods conceptualize explanations as *approximation models*, which are a type of simpler, human interpretable model that is created to reliably approximate the *functionality* of a more complex black box model. The approximation model itself is often and confusingly referred to as an explanation of the black box model. This approach contrasts with the treatment of 'explanation' in philosophy of science and epistemology, in which the term typically refers to *explanatory statements* that explain the causes of a given phenomenon (Mittelstadt et al. 2019).

The usage of 'explanation' in this fashion can be confusing. Approximation models are best thought of as tools from which explanatory statements about the original model can be derived (Mittelstadt et al. 2019). Explanatory statements themselves can be textual, quantitative, or visual, and report on several aspects of the model and its behaviours. Molnar

(2020: 25–26) proposes the following taxonomy of the types of outputs produced by interpretability methods:

- **feature summary statistic**: methods that return summary statistics indicating the strength of single features (e.g. feature importance) or groups of features (e.g. pairwise interaction strength);
- **feature summary visualization**: methods where summary statistics can also be visualized rather than listed quantitatively in a table. Visual outputs are preferable when reporting the partial dependence of a feature;
- **model internals**: methods where various aspects of the internals of a model can be reported, such as the learned weights of features, the learned structure of decision trees, or the visualization of feature detectors learned in convolutional neural networks;
- **data point**: methods were data points that help interpret a model can be reported, especially when working with textual or visual data. These data points can either exist in the model or be newly created to explain a particular output of the model. To be interpretable, any reported data points should ideally themselves be interpretable;
- **intrinsically interpretable model**: as discussed above, methods where globally or locally interpretable approximation models can be created to explain black box models. These models can then be further explained using any of the aforementioned methods and output types.

Further distinctions help to classify different types of explanations and interpretability methods. A basic distinction in interpretability can be drawn between *global* and *local interpretability*. This distinction refers to the scope of the model or outputs a given interpretability or explanatory method aims to make human comprehensible. Global methods

aim to explain the *functionality* of a model as a whole or across a particular set of outputs in terms of the significance of features, their dependencies or interactions, and their effect on outputs. In contrast, local methods can address, for example, the influence of specific areas of the input space or specific variables on one or more specific outputs of the model.

Models can be globally interpretable at a holistic or modular level (Molnar 2020). Holistic global interpretability refers to models which are comprehensible to a human observer in the sense that the observer can follow the entire logic or *functional* steps taken by the model which lead to all possible outcomes of the model (Guidotti et al. 2018). It should be possible for a single person to comprehend holistically interpretable models in their entirety (Lipton 2016). An observer would have 'a holistic view of its features and each of the learned components such as weights, other parameters, and structures' (Molnar 2020: 27).

Given the limitations of human comprehension and short-term memory, global holistic interpretability is currently only practically achievable on relatively simple models with few features, interactions, rules, or strong linearity and monotonicity (Guidotti et al. 2018). For more complex models, global interpretability at a *modular level* may be feasible. This type of interpretability involves understanding a particular characteristic or segment of the model, for example, the weights in a linear model or the splits and leaf node predictions in a decision tree (Molnar 2020).

With regards to local interpretability, a single output can be considered interpretable if the steps that led to it can be explained. Local interpretability does not strictly require that the entire series of steps be explained; rather, it can be sufficient to explain one or more aspects of the model that led to the output, such as a critically influential feature value (Molnar 2020; Wachter et al. 2018). A group of outputs is considered locally interpretable if the same methods to produce explanations of individual outputs can be applied to the group. Groups

can also be explained by methods that produce global interpretability at a modular level (Molnar 2020).

A further important distinction drawn in the literature concerns how and when interpretability is achieved in practice. Interpretability can be achieved by affecting the design and restricting the complexity of a model or by applying methods to analyse and explain the model after it has been trained (and deployed). Respectively, these can be referred to as *intrinsic interpretability* and *post hoc interpretability* (Lipton 2016; Molnar 2020; Montavon et al. 2017) or *reverse engineering* (Guidotti et al. 2018). Intrinsic interpretability can be further specified according to its target, which, according to Lipton (2016) and Lepri et al. (2017) can be a mechanistic understanding of the functioning of the model ('simulatability'), individual components ('decomposability'), or the training algorithm ('algorithmic transparency').

Interpretability methods

Development of methods for interpreting black box machine learning models has accelerated rapidly in recent years. While a full survey of methods remains beyond the scope of this chapter, the following taxonomy proposed by Guidotti et al. (2018) classifies methods according to the type of interpretability problem being solved:

- **model explanation methods**: these methods create a simpler, globally interpretable approximation model that acts as a global explanation of the black box model. These simplified models approximate the true criteria used to make decisions. Good approximations will reliably 'mimic the behavior of the black box' while remaining understandable to a target audience (Guidotti et al. 2018: 13). Such methods include 'single-tree approximations' which approximate the performance of the black box model in a single decision tree

(Craven and Shavlik 1996; Krishnan et al. 1999), 'rule extraction' methods which create human comprehensible decision rules that mimic the performance of the black box model (Andrews et al. 1995; Craven and Shavlik 1994), and varied global model-agnostic methods (Henelius et al. 2014; Lou et al. 2012, 2013);

- **outcome explanation methods**: these methods create a locally interpretable approximation model that can 'explain the prediction of the black box in understandable terms for humans for a specific instance or record' (Guidotti et al., 2018: 26). These methods do not need to be globally interpretable but rather only need to reliably explain 'the prediction on a specific input instance' (Guidotti et al. 2018: 13). Local approximations are accurate representations only of a specific domain or 'slice' of a model. As a result, there is necessarily a trade-off between the insightfulness of the approximated model, the simplicity of the presented function, and the size of the domain for which it is valid (Bastani et al. 2017; Lakkaraju et al. 2017). Such methods include saliency masks, which visually highlight areas of importance to an image classifier for a particular input class (Fong and Vedaldi 2017; Selvaraju et al. 2016), and varied local model-agnostic methods (Poulin et al. 2006; Ribeiro et al. 2016; Turner 2016);
- **model inspection methods**: these methods create a 'representation (visual or textual) for understanding some specific property of the black box model or of its predictions', such as the model's sensitivity to changes in the value of particular features or the components of the model that most influence one or more specific decisions (Guidotti et al. 2018: 14). As with outcome explanation methods, model inspection problems do not strictly require a

globally interpretable approximation to be created. Such methods include sensitivity analysis (Baehrens et al. 2010; Datta et al. 2016; Saltelli 2002), partial dependence plots (Adler et al. 2018; Hooker 2004; Krause et al. 2016), individual conditional expectation plots (Goldstein et al. 2015), activation maximization (Nguyen et al. 2016; Yosinski et al. 2015), and tree visualization (Thiagarajan et al. 2016).

- **transparent box design methods**: these methods produce a model that is locally or globally interpretable. This is not an approximation of a black box model but rather an original model (Guidotti et al. 2018). Rudin has influentially advocated for bypassing the problem of explanations by using interpretable models unless a significant and important loss in accuracy by failing to use a black box model can be demonstrated (Rudin 2019b). Methods commonly considered to be interpretable by design, given appropriate constraints on dimensionality or depth, include linear regression, logistic regression, regularized regression, and decision trees (Guidotti et al. 2018; Information Commissioner's Office, The Alan Turing Institute 2020; Molnar 2020). Other methods include rule extraction (Lakkaraju et al. 2016; Wang and Rudin 2015; Yin and Han 2003) and prototype and criticism selection (Bien and Tibshirani 2011; Fong and Vedaldi 2017; Kim et al. 2014).

This taxonomy does not capture the full range of interpretability methods. A class of methods not fully captured include what Lipton (Lipton 2016: 97) refers to as 'post-hoc interpretations' of specific behaviour. These include some of the methods classified as *outcome explanation methods*, such as visualizations (Simonyan et al. 2013; Tamagnini et al. 2017) and local model-agnostic explanations (Fong and Vedaldi 2017; Ribeiro et al. 2016), but also methods that create *user friendly verbal explanations*, including case-based

explanations (Caruana et al. 1999; Kim et al. 2014), natural language explanations (McAuley and Leskovec 2013), and counterfactual explanations (Wachter et al. 2018). Case-based explanation methods for non-case-based machine learning involve using the trained model as a distance metric to determine which cases in the training data set are most similar to the case or decision to be explained. Natural language explanations consist of text or visual aids describing the relationship between features of an input (e.g. words in a document) and the model's output (e.g. the classification of the document). Counterfactual explanations describe a dependency on external facts that led to a particular outcome or decision and a 'close possible world' in which a different, preferred outcome would have occurred (Wachter et al. 2018).

Interpretability methods can also be categorized according to their portability. Molnar (2020) distinguishes model-specific from model-agnostic methods, the latter of which can be applied to any type of machine learning model. Examples include dependence plots, feature interaction (Friedman and Popescu 2008; Greenwell et al. 2018; Hooker 2004), feature importance (Fisher et al. 2019), and local surrogates (Ribeiro et al. 2016). Example-based methods, which explain instances of the data set rather than groups of features or the model holistically, are also typically model-agnostic (Molnar 2020: 233). Examples include counterfactual explanations (Russell 2019; Wachter et al. 2018), adversarial examples (Goodfellow et al. 2014; Szegedy et al. 2013), prototypes and criticisms (Kim et al. 2016), influential instances (Koh and Liang 2017; Lundberg and Lee 2017), and case-based explanations (Caruana et al. 1999; Kim et al. 2014).

Transparency

A related but distinct topic often addressed in tandem with algorithmic interpretability is that of algorithmic transparency and accountability. Whereas interpretability has a *narrow* focus

of explaining the functionality or behaviour of an AI system or trained machine learning model, transparency and accountability have a *broad* focus on explaining the institutional and regulatory environment in which such systems are developed, deployed, and governed. In other words, interpretability is about understanding the system itself, whereas transparency and accountability are about understanding the people and organizations responsible for developing, using, and regulating it. Interpretability is often thought to be a key component of algorithmic transparency and accountability.

Given its broad aims, many approaches could conceivably be considered forms of algorithmic transparency and accountability. For our purposes, two broad categories can be distinguished: standardized documentation for training data sets and models, and impact assessments.

Standardized documentation

Standardized documentation refers to any method that prescribes a consistent form of disclosure about how AI systems and models are created, trained, and deployed in different decision-making contexts, services, and organizations. Many proposals for universal and sector-specific standards have been advanced in recent years, but none have yet been broadly adopted or tested.

Despite this, standardization initiatives frequently have common points of departure. The motivation for standardization in AI can be traced to comparable standards adopted in many industries describing the provenance, safety, and performance testing carried out on a product prior to release (Arnold et al. 2019). In this context, many initiatives are motivated by the usage, sharing, and aggregation of diverse data sets in AI, which runs the risk of introducing and reinforcing biases across different contexts of use (Bender and Friedman 2018; Gebru et al. 2018; Holland et al. 2018; Yang et al. 2018).

Data set documentation methods aim to help potential users of a data set to assess its appropriateness and limitations for training models for specific types of tasks. To do so, they generally require information about how the data sets are created and composed, including a list of features and sources of the data as well as information about how the data was collected, cleaned, and distributed (Gebru et al. 2018). Some approaches include disclosures and standardized statistical tests concerning ethical and legal considerations (Holland et al. 2018), including biases, known proxies for sensitive features (e.g. ethnicity, gender), and gaps in the data. Documenting such characteristics can help identify where problematic biases could be learned and reinforced by machine learning systems trained on the data which would otherwise remain unknown to developers and analysts (Gebru et al. 2018; Holland et al. 2018). As a secondary effect, standardized data set documentation may also drive better data collection practices (Holland et al. 2018) as well as consideration of contextual and methodological biases more generally.

Comparable initiatives exist for trained machine learning models. 'Model reporting' documentation is designed to accompany trained models when being deployed in contexts that differ from the training environment. For example, the 'model cards for model reporting' initiative calls for documentation describing various performance characteristics and intended contexts of use, including how performance changes when applied to different cultural, demographic, phenotypic, and intersectional (i.e. defined by multiple relevant attributes) groups (Mitchell et al. 2019). User-facing model documentation has also been proposed to enhance user trust and adoption. For example, 'FactSheets' have been proposed that would require a standardized declaration of conformity from AI suppliers addressing the purpose, performance, safety, security, and provenance of models in a user-friendly manner (Arnold et al. 2019). To complement data set and model documentation and pre-deployment testing

standards, toolkits have also been created to help identify and correct for biases in deployed models and AI systems (Bellamy et al. 2018).

Self-assessment frameworks

A second category of algorithmic transparency initiatives have created various self-assessment tools to help organizations evaluate AI systems at the point of procurement and deployment. These tools pose a series of questions to be answered by organizations in the procurement and deployment phase of AI. This approach builds on established types of legally required organizational disclosures in areas such as data protection, privacy, and environmental law (Article 29 Data Protection Working Party 2017; Mantelero 2018; Reisman et al. 2018). To date, self-assessment frameworks have largely been limited to public sector procurement of AI but in principle could be applied in the private sector as well.

'Algorithmic Impact Assessments' (AIA) are a prominent example of self-assessment frameworks (Reisman et al. 2018). The AIA developed by the AI Now Institute, for example, requires public agencies to consider four key elements prior to procurement: (a) potential impact on fairness, justice, bias, and similar concerns; (b) review processes for external researchers to track the system's impact over time; (c) public disclosure of the agencies' definition of 'automated decision system', current and proposed systems, and any completed self-assessments; and (d) solicitation of concerns and questions from the public. AI Now has also called on governments to establish enhanced due process mechanisms to support individual and community redress (Reisman et al. 2018). The AIA framework has since been implemented and adapted by the Canadian government to govern procurement of automated decision-making systems across the public sector (Government of Canada 2020). The European Commission has also recently developed a 'Trustworthy AI Assessment List' that is operationally similar to an AIA (High Level Expert Group on Artificial Intelligence 2019).

The list poses a series of questions on topics such as fundamental rights; human agency and oversight; technical robustness; and safety, diversity, and accountability.

Philosophy of explanations

Examining prior work on AI interpretability and transparency in the context of prior work on explanations in the philosophy of science can be useful to identify the field's major trends, gaps, key open questions, and potentially their answers. Explanations of scientific and everyday phenomena have long been studied in the philosophy of science. Explanations, and more broadly epistemology, causality, and justification, have been the focus of philosophy for millennia, making a complete overview of the field unfeasible. What follows is a brief overview of key distinctions and terminology relevant to surveying interpretability and transparency in AI.

While much variation and debate can be observed in prior work (Ruben 2004; Salmon 2006), an explanation of a given phenomenon is usually said to consist of two parts:

- the *explanadum* or a sentence describing the phenomenon to be explained (Hempel and Oppenheim 1948: 136–137). The phenomenon can be of any level of specificity from a particular fact or event, such as a particular decision produced by a model, to general scientific laws or holistic descriptions of a model;
- the *explanans* or the sentences which are thought to explain the phenomenon (Hempel and Oppenheim 1948: 136–137). Depending upon the type of explanation, audience, and specific questions asked, the explanans can be as simple as a single sentence or as complex as a full causal model.

In the philosophy of science, much work is dedicated to theories of *scientific explanation*. According to this tradition, 'explanatory knowledge is knowledge of the causal mechanisms,

and mechanisms of other types perhaps, that produce the phenomena with which we are concerned' (Salmon 2006: 128). A related notion, causal explanation, refers to a type of explanation of an event that provides 'some information about its causal history' (Lewis 1986: 217–218). Within this tradition, a complete or scientific explanation would consist of a set of *explanans* describing the full causal history of a phenomenon (Hempel 1965; Ruben 2004; Salmon 2006). This type of scientific explanation will involve general scientific relationships or universal laws and can be considered an idealized form of explanation of the sort pursued but rarely obtained through scientific investigation (Hempel 1965).

As this definition of an ideal scientific explanation suggests, explanations can be classified in terms of their *completeness* or the degree to which the entire causal chain and necessity of an event can be explained (Ruben 2004). Completeness can be used to distinguish *scientific* and *everyday* explanations (Miller 2019) or *full* and *partial* causal explanations (Ruben 2004), each of which addresses the causes of an event but to different degrees of completeness. Everyday explanations of the type typically requested in daily life address 'why particular facts (events, properties, decisions, etc.) occurred' rather than general scientific relationships (Miller 2019: 3).

The terminology is not, however, consistent across theories of explanation. As Ruben (2004: 19) notes:

> Different theories disagree about what counts as a full explanation. Some will hold that explanations, as given in the ordinary way, are full explanations in their own right; others (like Hempel) will argue that full explanations are only those which meet some ideal, rarely if ever achieved in practice. A partial explanation is simply a full explanation (whatever that is) with some part of it left out. On any theory of explanation, we sometimes do not say all that we should say if we were explaining in full. Sometimes we assume that the

> audience is in possession of facts which do not stand in need of repetition. At other times, our ignorance does not allow us to fill some of the explanatory gaps that we admit occur. In such cases, in which we omit information for pragmatic or epistemic reasons, we give partial explanations.

Most of the methods discussed in this chapter can be considered partial or everyday explanations. These methods report a selection of the total set of causes of a phenomenon or create a simplified approximation of a more complex phenomenon to make it humanly comprehensible. Both examples are partial because they do not report the full set of causes of a phenomenon, for example, the full causal chain of collecting and cleaning training and test data or the causes of the phenomena reported in this data.

Full or scientific explanations can nonetheless serve as an idealized endpoint for global interpretability in AI. If a user asks how a model was trained, a good explanation would resemble a full causal explanation, only limited to the internals of the model (e.g. feature values and interdependencies) and training algorithm rather than universal laws. Similarly, global explanations of model *functionality* will necessarily contain causal information concerning, for example, dependencies between features. Scientific explanations traditionally conceived are also relevant to the need for interpretable machine learning models in research, especially on causality and inference (Pearl 2019; Schölkopf 2019).

Recent decades have seen an increase in the attention paid to theories of contrastive explanations and counterfactual causality (Kment 2006; Lewis 1973; Pearl 2000; Woodward and Zalta 2003). Contrastive theories suggest that causal explanations inevitably involve appeal to a counterfactual case, be it a cause or an event, which did not occur. Woodward (2005: 6) describes these types of explanations as answers to 'what-if-things-had-been-different' questions. A canonical example is provided by Lipton (1990: 256):

> To explain why P rather than Q, we must cite a causal difference between P and not-Q, consisting of a cause of P and the absence of a corresponding event in the history of not-Q.

Contrastive theories of explanation are, of course, not without criticism. Ruben (2004), for example, has suggested that, even if causal explanations are inevitably contrastive in nature (which he doubts), this feature can be accommodated by traditional theories of explanation. Regardless of one's position on this debate, contrastive explanations remain interesting for AI because they address a particular event or case and would thus appear to be simpler to create than global explanations of model *functionality* (Mittelstadt et al. 2019; Wachter et al. 2018).

Other types of explanations beyond scientific explanations exist which are relevant to interpretability in AI. As Hempel writes, 'explaining the rules of a contest, explaining the meaning of a cuneiform inscription or of a complex legal clause or of a passage in a symbolist poem, explaining how to bake a Sacher torte or how to repair a radio' are all uses of the term 'explain' which do not involve causal, scientific explanations (Hempel 1965: 412–413). In these cases, explanations can be given that do not have a clear dependence on universal or scientific laws.

These observations are relevant to the question of interpretability in AI insofar as it indicates that 'explanation' is not a singular concept but rather a catch-all for many different types of interlocutory acts. A person impacted by an AI system (e.g. a criminal risk scoring system) could ask why they were classified as high risk but equally how the model was trained, on which data, and why its usage (and design) is morally or legally justified. Work on interpretability in AI, as well as regulatory interest in the subject, similarly reflects that explanations of AI are being requested in connection to a particular entity, be it a specific decision, event, trained model, or application (Mittelstadt et al. 2019). The explanations requested are thus not full scientific explanations as they need not appeal to general

relationships or scientific laws but rather, at most, to causal relationships between the set of variables in a given model (Woodward 1997). Rather, what is being requested are everyday explanations either of how a trained model functions in general or of how it behaved in a particular case.

For the purposes of this chapter, I will primarily discuss methods for producing explanations in AI systems that answer *functional* questions, such as how a model functions globally and locally, or how a particular classification was reached. Such explanations, while primarily technical answers to 'Why?' questions (e.g. why I was classified as 'high risk'), simultaneously provide essential information to answer related questions concerning the accuracy, reliability, safety, fairness, bias, and other aspects of the system.

A further distinction can be drawn between explanation as a process or act and explanation as a product of that act. This linguistic feature is known as process–product ambiguity. In philosophy of science, much work has been dedicated to explanation as both a product and a process and their dependency (if any) (Hempel 1965; Ruben 2004). As a product, the question being asked is essentially 'What information has to be conveyed in order to have explained something?' Explanations as products can be classified and described according to the type of information they convey (Ruben 2004). As a process, the act of explaining and the intention of the explainer are thought to influence the information conveyed by an explanation (Achinstein 1983). The explanation as a product is thus 'an ordered pair, in part consisting of a proposition, but also including an explaining act type' (Ruben 2004: 8).

The process and product accounts of explanation are incompatible (Ruben 2004); however, for the purposes of this chapter, an answer to the process–product ambiguity is not needed. Rather, the distinction reveals that in designing explanations and explanatory methods for AI, attention must be given not only to *what* information the explanation

contains but also to *how* this information is conveyed to its audience. This distinction between the *what* and *how* of explanations in AI is key to evaluating the relative utility of the methods discussed above and the quality of different types of explanations.

Evaluating the quality of explanations

Prior philosophy work on explanations provides a robust foundation to explore how the quality of different types of explanations and approximations of AI *functionality* or *behaviour* can be evaluated. Within the philosophy of science, explanatory *pragmatists* suggest that 'explanation is an interest-relative notion . . . explanation has to be partly a pragmatic concept' (Putnam 1978: 41). In other words, the requirements for a full explanation will vary according to the needs and interests of the audience.

This approach is a departure from causal theorists, who draw a clear distinction between the ideal of a full explanation and the pragmatics of giving a good explanation. The former is a question of the information the explanatory product must contain, while the latter is a question of how parts of that information, or a partial explanation, is crafted and communicated to an audience according to their particular interests and requirements. According to Ruben (2004: 22),

> . . . how we select from the full list of explanatory relevant features in order to obtain the ones required in a particular (partial) explanation we may offer is a pragmatic and audience-variant question. A partial explanation is one that omits certain relevant factors; a full explanation is one that includes all relevant factors . . . A partial explanation may be good relative to one set of circumstances, but bad relative to another, in which interests, beliefs, or whatever differ.

The question is thus whether, as in the deductive-nomological approach (Salmon 2006), the ideal of a full explanation exists independently of the concept of a good explanation. The imagined 'ideal explanatory text' provides a benchmark for complete scientific explanations (Salmon 2006). For explanatory pragmatists, this distinction collapses. Once collapsed, context becomes an essential determinant of a good explanation. Whereas traditionalists conceive of the concept of explanation as 'a relation like description: a relation between a theory and a fact', pragmatists view it as 'a three-term relation between theory, fact, and context' (Fraassen 1980: 156).

According to pragmatists, good explanations exceed merely correct explanations by being aligned with the needs, interests, and expertise of the agents requesting the explanation (Achinstein 2010; Lewis 1986). It follows that a universal ideal of a best possible explanation of any given phenomenon does not exist; while an ideal correct explanation is possible, what makes an explanation good (or 'the best') is dependent upon the context and audience to which it is given (Achinstein 2010).

Regardless of one's position as a traditionalist or pragmatist, a distinction can be drawn between the truth or correctness of an explanation and how successful that explanation is at communicating relevant information to a given audience (Ruben 2004). Achinstein (2010) describes this as the distinction between *correct* explanations and *good* explanations. A full scientific explanation can be correct insofar as the causes it attributes to a phenomenon are truthful or valid and yet be a bad explanation when evaluated as an act of communication, for example, because the information conveyed is so complex as to be incomprehensible to the recipient. Similarly, an explanation may also be considered inadequate not because the information communicated is false but because it is incomplete or inadequate to answer the question posed or the needs of a specific audience (Lewis 1986; Putnam 1978).

Compared to the distinctions drawn above, this is a subtle but important difference. In evaluating an explanation in AI, we can distinguish quality in terms of *causal validity*, or its truthfulness and completeness, and quality in terms of *meaningfulness* or how effective it is at conveying a relevant set of information to a given audience. This distinction holds across both traditional and pragmatic schools of thought, which differ only on whether the meaningfulness of an explanation should be considered a quality of the explanation itself (as pragmatists do) or a quality of the act of selecting and communicating the explanation (as traditionalists do). For the purpose of this chapter, selecting a particular school of thought is unnecessary so long as the distinction between causal validity and meaningfulness is recognized.

Characteristics of 'good' explanatory products

Building on this distinction between validity and meaningfulness, many characteristics have been proposed in the field of AI interpretability to evaluate the quality of explanations and approximations. Following the preceding discussion (see the section 'Philosophy of explanations'), a further distinction can be drawn between the quality of the *explanans* itself and the quality of the process by which the *explanans* is communicated to the explainee. What follows is an overview of characteristics for producing and communicating high-quality explanations that have been proposed in literature on AI interpretability as well as empirical work describing how humans give and receive explanations in psychology and cognitive science (Miller 2019). We begin with characteristics to evaluate the quality of explanatory products.

Contrastive

Based on a representative review of empirical evidence in psychology and cognitive science, Miller (2019: 3) argues that good everyday explanations are contrastive insofar as explanations are 'sought in response to particular counterfactual cases . . . That is, people do not ask why event P happened, but rather why event P happened instead of some event Q.' Based on the reviewed evidence, Miller found that people psychologically prefer contrastive explanations. Further, this preference cannot be reduced solely to the relative simplicity of contrastive explanations against full causal explanations (Miller 2019: 28). In AI, best practices for computing contrastive explanations will be specific to context, application, or user because a comparison point or preferred alternative outcome must be identified (Molnar 2020; Wachter et al. 2018).

Abnormality

'Normal' behaviour is thought to be 'more explainable than abnormal behaviour' (Miller 2019: 41). The perceived abnormality of an event has thus been found to drive the preference for contrastive explanations in practice that can explain why a normal or expected event did not occur (Gregor and Benbasat 1999; Hilton and Slugoski 1986; McClure et al. 2003; Molnar 2020; Samland and Waldmann 2014). Many characteristics of AI behaviour can set it apart as abnormal. Lim and Dey (2009), for example, found a positive relationship between the perceived 'inappropriateness' of application behaviour and user requests for contrastive explanations. Violation of ethical and social norms can likewise set an event apart as abnormal (Hilton 1996). The practical importance of abnormality for good everyday explanations suggests that explanations of AI behaviour should describe input features that are 'abnormal in any sense (like a rare category of a categorical feature)' if they influenced the behaviour or outcome in question (Molnar 2020: 32).

Selectivity

Full scientific explanations are rarely if ever realized in practice. Multiple correct but incomplete explanations are normally possible that list different causes for the *explanadum*. A given cause may be incomplete insofar as it is not the sole cause of the event but may nonetheless convey useful information to the explainee (Ylikoski 2013). As Miller (2019: 3) argues, 'Explanations are selected—people rarely, if ever, expect an explanation that consists of an actual and complete cause of an event. Humans are adept at selecting one or two causes from a sometimes infinite number of causes to be the explanation.' Selection involves choosing the most relevant set of causes for a given phenomenon and disregarding other less relevant but valid causes on the basis of local requirements. Selection is necessary to reduce long causal chains to a cognitively manageable size (Hilton 1996).

For AI explanations, selection means choosing key features or evidence to be emphasized in an explanation or user interface based, for example, on their relative weight or influence on a given prediction or output (Biran and McKeown 2014; Poulin et al. 2006) and the explainee's subjective interests and expectations (see the section 'Characteristics of 'good' explanatory processes'). To facilitate selection of relevant explanans from the overall possible set of valid explanans, good explanations should clearly communicate the degree of importance or influence of a given feature or set of features on the instance or outcome being explained (Molnar 2020).

Complexity and sparsity

The need for selectivity in explaining AI behaviours to meet local requirements points to the need to conceptualize and measure the relative complexity of different possible valid explanations (Achinstein 1983; Miller 2019). Many metrics exist to evaluate explanation

complexity relative to a target model or set of outputs. Complexity can be defined in relation to a model's size, such as the number and length of rules, features, or branches in a decision tree (Deng 2019; Guidotti et al. 2018; Rudin 2019b), or in terms of the linearity and monotonicity of the relationships between variables (Guidotti et al. 2018). Alternatively, complexity can be defined according to sparsity or the number of explanatory statements given to explain a black box model or specific output as well as the number of features and interactions addressed in these statements.

In AI, sparse explanations or approximation models are those which have low dimensionality or address a small number of features and interactions. Good sparse explanations are those which include a cognitively manageable set of highly relevant causes or statements according to the explainee's interests and expertise (Molnar 2020; Russell 2019). Methods such as case-based explanations and counterfactual explanations that provide a sparse explanation of changes necessary to reach a different, preferred outcome can bypass the difficult challenge of explaining the internal state of trained models to a significant extent (Caruana et al. 1999; Wachter et al. 2018). Approximation methods can also help but must grapple with a three-way trade-off between the approximation's fidelity, comprehensibility, and domain size (see the section 'Interpretability').

Novelty and truthfulness

A set of closely related characteristics common to theories of explanation concerns the novelty and truthfulness of the explanans. Good explanations should be novel, meaning they do not merely repeat information about the explanadum that is already known by the explainee but rather provide new, unknown information that helps explain the explanadum (Molnar 2020; Salmon 2006). To be informative, explanations should not be entirely reducible to presuppositions or beliefs that the recipient of the explanation already holds

(Hesslow 1988). In AI, novelty can be measured, for example, in the extent to which an explanation reflects whether the instance being explained 'comes from a "new" region far removed from the distribution of training data' (Molnar 2020: 28).

Good explanations should likewise be truthful, meaning the statements contained in the explanans should be accurate or correct. In AI explanations, the dependencies between variables described or the causes attributed to an outcome should be correct (Molnar 2020; Russell 2019). Simply put, the more accurate the explanation, the better it is at enhancing the explainee's understanding of the explanadum. In practice, accuracy can be measured, for example, in terms of the performance of the explanation in predicting future behaviours based on unseen input data (Molnar 2020).

Representativeness, fidelity, consistency, and stability

A final set of characteristics addresses the intra-model and inter-explanation performance of an explanation or approximation. For explanations of more than a single output or group of outputs, representativeness is a key characteristic. Global or local approximations can be evaluated in terms of their representativeness of outputs or instances in the model reliably explained by the approximation (Molnar 2020). As a rule of thumb, the quality of an approximation increases based on the number of instances or outputs of the model it can reliably and accurately explain.

Fidelity is closely related to representativeness insofar as the latter is implicitly linked to the accuracy of the explanation over multiple insurances. Fidelity refers to the performance of the approximation against the black box model; approximations with high fidelity will approximate the performance of the black box model as closely as possible, including accurate predictions as well as errors.

The consistency of the approximation is also relevant in this context, which can be measured in terms of performance of the approximation across different black box models that have been trained on the same data to perform the same task. Stability performs a similar role to consistency. They differ in that stability is concerned with comparing the performance of explanations for 'similar instances for a fixed model'. Stable explanations will not substantially change when explaining a set of instances that only have slight variation in feature values (Molnar 2020).

Characteristics of 'good' explanatory processes

Following the distinction between causal validity and meaningfulness, the quality of explanations is dependent not solely on the content of the explanation but also on how this content is tailored to the explainee and communicated in practice. Many factors of good explanatory processes, such as the appropriate complexity and scope of explanations provided, are dependent on the context in which an AI system is used. The following characteristics of good explanatory processes have been proposed in the literature.

Interactivity and usability

Giving an explanation is a social communicative act based on interaction and information exchange between one or more explainers and explainees (Slugoski et al. 1993). Information exchange occurs through dialogue, visual representation, or other means (Hilton 1990). In AI, giving an explanation should be viewed not as a one-way exchange of information but rather as an interactive process involving a mix of human and AI agents.

Further, explanations are iterative insofar as they must be selected and evaluated on the basis of shared presuppositions and beliefs. Iteration may be required to communicate effectively or clarify points of confusion on the path towards a mutually understood

explanation. While a given output can have many causes or important features, explainees will often only be interested in a small subset of these that are relevant to a specific question or contrastive case. It is the task of the explainer to select explanans from this subset of all possible causes or features. The chosen explanans may not satisfy the requirements of the explainee, requiring subsequent questioning and the generation of a new, more relevant explanans (Miller 2019: 4).

The quality of explanatory processes in AI can therefore be evaluated in terms of the quality of the interaction and iteration between explainee and explainer. Forms of explanation that are interactive and can help the explainee interrogate the model to accomplish specific tasks of interest are seen as better than explanations which consist of standardized or fixed content (Guidotti et al. 2018: 7). Explanations of AI *functionality* or *behaviour* can be given both by human workers tasked with explaining the system to affected parties and potentially by the AI system itself, for example, through an interpretability interface (Kayande et al. 2009; Martens and Provost 2013; Mittelstadt et al. 2019; Wexler et al. 2019).

Local relevance

As the preceding characteristics indicate, good explanations should be tailored towards the relative interests of the explainee (Miller 2019; Molnar 2020). They should answer, or help to answer, questions of interest to their audience (Miller 2019; Slugoski et al. 1993). Software engineers, regulators, deploying institutions, end-users, and other people request explanations for different reasons and seek answers to different questions (Miller 2019; Mittelstadt et al. 2019). Explanations that are not tailored to answer the specific question(s) being asked may fail to communicate relevant information to their audience.

Local comprehensibility

For explanations to succeed in communicating relevant information to the explainee, they must also be comprehensible to their recipient. Local comprehensibility refers to the degree to which explanations communicate information at a scope and level of complexity that matches the audience's expertise (Molnar 2020). Explanations that include all factors that led to a particular prediction or behaviour can be correct and complete explanations and yet incomprehensible, depending on their audience. For example, complete explanations may be useful for purposes of debugging a system or to meet legal requirements (Molnar 2020: 35) but useless to a user trying to understand which factors of their financial history most influenced the outcome of their loan application (Wachter et al. 2018).

Local comprehensibility and relevance are intrinsically linked. Software engineers, for example, may prefer more accurate but opaque models or more complete but complex explanations to help debug and refine their system, whereas end-users interested in the key reasons for a given decision may prefer explanations that are simpler or narrower in scope (Mittelstadt et al. 2019; Wachter et al. 2018). Urgency is similarly important; time-sensitive requests can necessitate simpler but incomplete explanations (Guidotti et al. 2018).

Overall, good explanatory processes in AI should be sensitive to the reason an explanation is being requested as well as the motivation and local needs of the explainee. In this regard, standardized forms of disclosure, including many of the transparency frameworks describe above (see the section 'Transparency'), can fail as good explanations if they are not adapted for different audiences.

Open challenges in AI interpretability and transparency

As the preceding discussion indicates, there are many open questions when it comes to designing effective products and processes to explain the functionality and behaviour of AI

systems. To conclude, I consider three key open challenges facing the field of AI interpretability and transparency concerning the development of common standards for (a) 'good' explanations; (b) deterring deception through explanations; and (c) consistent and practically useful transparency frameworks.

Common standards for 'good' explanations

As discussed in this chapter, many methods have been developed to explain how autonomous systems function both generally and for specific decisions. However, while many approaches exist, the adoption of common standards for 'good' explanations that enhance the useability of autonomous systems remain nascent. To define such standards, we need to understand what makes an explanation informative and effective in practice. Empirical research into the local effectiveness and acceptability of different interpretability and transparency methods for different types of AI applications is urgently needed.

To date, a majority of work on AI interpretability has addressed methods for creating global and local approximations of black box models. While useful for purposes of testing and debugging black box models, the utility of these approaches for explaining model behaviour are less clear (Ribeiro et al. 2016; Selvaraju et al. 2016; Simonyan et al. 2013). In particular, it is unclear how useful such simplified human comprehensible approximations are for non-experts. Local approximations in particular 'can produce widely varying estimates of the importance of variables even in simple scenarios such as the single variable case, making it extremely difficult to reason about how a function varies as the inputs change' (Wachter et al. 2018: 851). Conveying these limitations in a consistent and reliable way to experts and non-experts alike remains nascent, which raises questions over their utility for answering questions about specific model behaviour.

It is in this context that Box's maxim, 'All models are wrong, but some are useful' (Box 1979) is illuminating. Treating local approximations as explanations of model behaviour would suggest that they provide reliable knowledge of how a complex model functions, but this has yet to be proven in practice across different types of AI applications and interpretability methods. For approximation models to be trusted, explainees must understand the domain over which the approximation is 'reliable and accurate, where it breaks down, and where its behaviour is uncertain' (Hesse 1965; Mittelstadt et al. 2019: 3). Without this information, approximations will be at best poorly comprehensible and at worst misleading because they are often inaccurate or unreliable outside a specific domain or set of instances (Mittelstadt et al. 2019).

Local approximations face difficulties with generalizability, arbitrariness in choice of domain, and the potential to mislead recipients unless the domain and epistemic limitations of the approximation are known (Mittelstadt et al. 2019). Standards for 'good' approximations are thus urgently needed that require information regarding their limitations to be clearly documented and communicated when an approximation is offered as an explanation of a black box model. To date, little work has been done on testing and validating approximations in real-world scenarios; going forward, this critical gap in AI interpretability needs to be closed.

Deception in explanations

The relative lack of research and methods to test and evaluate the veracity, objectivity, and overall quality of explanations and approximation models is concerning as even an accurate explanation can be used to inform or handcrafted to mislead (Lakkaraju and Bastani 2019). Features or causes can be intentionally selected in explanations to convey information according to the controller's preferences and to hide potentially worrying factors. For

example, a single explanation of law school admissions could highlight the classifier's dependency on ethnicity, entrance exam results, or grade point average over several years (Russell 2019).

The type of explanation provided can influence the explainee's opinion on the importance of features in an output or classification (Lombrozo 2009; Poulin et al. 2006). By explicitly shaping the choice of domain and the choice of approximation, it is possible to distort how the importance of variables are reported, to alter whether they are claimed to positively or negatively influence decisions, or to eliminate the correlation between them. This selectiveness grants systems controllers the power to alter people's beliefs about the reasons for a system's behaviour and to instil undue confidence in the system's performance and trustworthiness. Understanding how and why a particular explanation was chosen by the explainer is particularly important for the selection of contrastive cases for contrastive explanations.

The act of giving an explanation is not neutral (Mittelstadt et al. 2019). Some scholars, for example, have suggested that explanation-giving is not primarily directed to the truth but aims at persuasion (Mercier and Sperber 2011). Agents seeking to be perceived as trustworthy have an incentive not merely to explain their behaviour as accurately as possible but also to provide explanations that persuade other agents to perceive them as trustworthy. This incentive is seemingly at odds with the push to adopt AI systems for the sake of accuracy or efficiency. Suboptimal but simpler actions can improve transparency and communication between institutions, users, and end-users but can make systems and institutions less trustworthy and degrade the resiliency of the trust relationship if end-users experience poor outcomes (Glikson and Woolley 2020). Promoting interpretability or transparency can create incentives for systems to prefer actions for which there are easier or simpler explanations but which may not be optimal for the user (Mercier and Sperber 2011).

There is a clear and immediate risk of malicious actors using explanations not to inform but to mislead. Ethically or legally significant influences on a decision (e.g. sensitive features such as ethnicity) could be hidden from explainees interested in the legality of university admissions. Explainees can be subtly nudged by the choice of explanans to adopt a preferred belief or take a preferred action of the explainer, for example, not contesting admissions outcomes on the grounds of discrimination. Solutions to mitigate the risk of deception via explanation are urgently needed if AI interpretability and transparency are to make AI systems more accountable and trustworthy.

The effectiveness of self-assessment transparency frameworks

Self-assessment frameworks are intended to enhance the traceability of AI systems by improving organizational accountability, helping to identify potential ethically problematic impacts, and providing a starting point for redress for affected individuals and communities. If successful, each of these effects could enhance public trust and acceptance of AI systems.

However, while recognizing their potential utility, self-assessment frameworks have a number of inherent weaknesses. To be effective, self-assessment must be timely, transparent, honest, and critical. Organizations must invest the resources necessary to train staff to critically assess internal procurement procedures and the (potential) external impact of the system. Critical analysis requires an organizational culture that rewards honest assessment. Even if staff are well trained and rewarded for their honesty, further investment is needed to make self-assessment more than a ‘one off’ occurrence. The impact of AI systems cannot be perfectly predicted prior to deployment; unexpected and novel effects can emerge over time that, by definition, can only be captured through iterative self-assessment (Mittelstadt et al. 2016). To assess impact over time, internal procedures must be established to record system behaviour longitudinally. Sustaining the quality of such procedures over time has historically

proven difficult, with comparable self-assessment frameworks in other domains effectively becoming empty ‘checklists’ over time (Manders-Huits and Zimmer 2009; Mittelstadt 2019; Põder and Lukki 2011). Assuming these elements are in place, decisions must still be made about what, when, and how to involve external researchers and the public in the assessment process and how to publicly disclose results. Self-assessments that are perceived as incomplete, dishonest, inaccessible, or otherwise faulty will not have their intended effect on public trust.[2]

In short, self-assessment frameworks are not guaranteed to be effective without significant organizational commitment to staff training, organizational culture, sustainable and critical assessment procedures, public and researcher involvement, and open disclosure of results. These requirements cannot be guaranteed by developing universal guidelines or procedures for self-assessment because the potential impact of AI systems varies greatly according to context and application type (High Level Expert Group on Artificial Intelligence 2019; Mittelstadt 2019).

Conclusion

This chapter has reviewed the key concepts, approaches, difficulties, literature, and overall state of the art in interpretability and transparency in AI. Numerous methods to provide explanations, approximations, and standardized disclosures for the development, *functionality*, or *behaviour* of AI systems have been reviewed. To evaluate the quality of emergent approaches in the field, lessons can be learned from prior work on explanations in the philosophy of science. Explanations as products can be evaluated in terms of their contrastiveness, abnormality, selectivity, complexity and sparsity, novelty and accuracy, representativeness, fidelity, and consistency. Likewise, the act of giving an explanation can

be critiqued in terms of its interactivity and usability and its relevance and comprehensibility to local stakeholders.

These characteristics of explanations as products and processes point towards a clear conclusion: interpretability and transparency in AI cannot possibly be achieved through a 'one-size-fits-all' approach. Different audiences, models, behaviours, and use cases will demand different forms of explanations. And yet, despite this diversity of products and methods, common ground exists to evaluate the quality of explanations in AI.

Critical open challenges, of course, remain. Consistent frameworks to evaluate explanations of AI must first be widely adopted for common standards of 'good' explanations to be enforceable through ethical or regulatory means. Likewise, vigilance is required to ensure that explanations and transparency mechanisms are used honestly and accurately and never to deceive or mislead. If AI systems are to deliver on their promise of more accurate, efficient, and accountable decision-making, solving these challenges and implementing common standards for interpretability and transparency is essential.

Author's note

Sections of this chapter are adapted from: Mittelstadt, B., Russell, C., and Wachter, S. (2019), 'Explaining Explanations in AI', *Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19*, 279–288. doi: https://doi.org/10.1145/3287560.3287574. This work has been supported by research funding provided by the British Academy grant no. PF\170151, Luminate Group, and the Miami Foundation.

References

Achinstein, Peter. (1983), The Nature of Explanation (Oxford: Oxford University Press on Demand).

Achinstein, Peter. (2010), Evidence, Explanation, and Realism: Essays in Philosophy of Science (Oxford: Oxford University Press).

Adler, Philip, Falk, Casey, Friedler, Sorelle A., Nix, Tionney, Rybeck, Gabriel, Scheidegger, Carlos, Smith, Brandon, and Venkatasubramanian, Suresh (2018), 'Auditing Black-Box Models for Indirect Influence', Knowledge and Information Systems 54, 95–122.

Andrews, Robert, Diederich, Joachim, and Tickle, Alan B. (1995), 'Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks', Knowledge-Based Systems 8, 373–389.

Arnold, Matthew, Bellamy, Rachel K., Hind, Michael, Houde, Stephanie, Mehta, Sameep, Mojsilović, Aleksandra, Nair, Ravi, Ramamurthy, Karthikeyan Natesan, Olteanu, Alexandra, Piorkowski, David, Reimer, Darrell, Richards, John, Tsay, Jason, and Varshney, Kuhan R. (2019), 'FactSheets: Increasing Trust in AI Services through Supplier's Declarations of Conformity', IBM Journal of Research and Development 63, 6:1–6:13. doi: https://doi.org/10.1147/JRD.2019.2942288.

Article 29 Data Protection Working Party (2017), Guidelines on Data Protection Impact Assessment (DPIA) and Determining Whether Processing is 'Likely to Result in a High Risk' for the Purposes of Regulation 2016/679, 17/EN WP 248.

Baehrens, David, Schroeter, Timon, Harmeling, Stefan, Kawanabe, Motoaki, Hansen, Katja, and Müller, Klaus-Robert. (2010), 'How to Explain Individual Classification Decisions', Journal of Machine Learning Research 11, 1803–1831.

Bastani, Osbert, Kim, Carolyn, and Bastani, Hamsa (2017), 'Interpretability via Model Extraction', arXiv Preprint, 1706.09773 [cs, stat].

Bellamy, Rrachel K.E., Dey, Kuntal, Hind, Michael, Hoffman, Samuel C., Houde, Stephanie, Kannan, Kalapriya, Lohia, Pranay, Martino, Jacquelyn, Mehta, Sameep, Mojsilovic, Aleksandra, Nagar, Seema, Ramamurthy, Karthikeyan N., Richards, John, Saha, Diptikalyan, Sattigeri, Prasanna, Singh, Moninder, Varshney, Kush R., and Zhang, Yunfeng. (2018), 'AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias', arXiv Preprint, 1810.01943 [cs].

Bender, Emily M., and Batya Friedman. (2018), 'Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science', Transactions of the Association for Computational Linguistics 6,: 587-604. doi: https://doi.org/10.1162/tacl_a_00041.

Berendt, Bettina, and Sören Preibusch. (2017), 'Toward Accountable Discrimination-Aware Data Mining: The Importance of Keeping the Human in the Loop—and under the Looking Glass', Big Data 5, 135–152. doi: https://doi.org/10.1089/big.2016.0055.

Bien, Jacob, and Robert Tibshirani. (2011), 'Prototype Selection for Interpretable Classification', Annals of Applied Statistics 5, 2403–2424. doi: https://doi.org/10.2307/23069335.

Biran, Or, and Kathleen McKeown. (2014), 'Justification Narratives for Individual Classifications', in Proceedings of the AutoML Workshop at ICML, 1-7.

Box, George EP. (1979), 'Robustness in the Strategy of Scientific Model Building', in Robert L. Launer and Graham N. Wilkinson, ed., Robustness in Statistics (Elsevier: New York), 201–236.

Burrell, Jenna. (2016), 'How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms', Big Data & Society 3. doi: https://doi.org/10.1177/2053951715622512

Caruana, Rich, Kangarloo, Hooshang, Dionisio, John David, Sinha, Usha, and Johnson, David. (1999), 'Case-Based Explanation of Non-Case-Based Learning Methods', Proceedings of the AMIA Symposium, 212–215.

Citron, Danielle Keats, and Frank Pasquale. (2014), 'The Scored Society: Due Process for Automated Predictions', Washington Law Review 89, 1.

Craven, Mark, and Jude W. Shavlik. (1994), 'Using Sampling and Queries to Extract Rules from Trained Neural Networks', in William W. Cohen and Haym Hirsh, eds, Machine Learning Proceedings 1994 (San Francisco, CA: Morgan Kaufmann), 37–45. doi: https://doi.org/10.1016/B978-1-55860-335-6.50013-1.

Craven, Mark W., and Jude W. Shavlik. (1995), 'Extracting Tree-Structured Representations of Trained Networks', in Proceedings of the 8th International Conference on Neural Information Processing Systems (Cambridge: MIT Press), 24–30.

Datta, Anupam, Shayak Sen, and Yair Zick. (2016), 'Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems', Institute of Electrical and Electronics Engineers, 598–617. doi: https://doi.org/10.1109/SP.2016.42.

Deng, Houtao. (2019), 'Interpreting Tree Ensembles with Intrees', International Journal of Data Science and Analytics 7, 277–287.

Doshi-Velez, Finale, and Been Kim. (2017), 'Towards a Rigorous Science of Interpretable Machine Learning', arXiv Preprint, 1702.08608 [cs, stat].

Fisher, Aaron, Cynthia Rudin, and Francesca Dominici. (2019), 'All Models Are Wrong, But Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously', Journal of Machine Learning Research 20, 1–81.

Fong, Ruth C., and Andrea Vedaldi. (2017), 'Interpretable Explanations of Black Boxes by Meaningful Perturbation', arXiv Preprint, 1704.03296.

Fraassen, Bas C. van (1980), The Scientific Image (Oxford: Clarendon Press).

Franzke, Aline Shakti, Bechmann, Anja, Zimmer, Michael, and Ess, Charles M. (2020), Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers.

Friedman, Jerome H., and Popescu, Bogdan E. (2008), 'Predictive Learning via Rule Ensembles', Annals of Applied Statistics 2, 916–954.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. (2018), 'Datasheets for Datasets', arXiv Preprint, 1803.09010.

Glikson, Ella, and Anita Williams Woolley. (2020), Human Trust in Artificial Intelligence: Review of Empirical Research, Academy of Management Annals 14, 627-660.

Goldstein, Alex, Adam Kapelner, Justin Bleich, and Emil Pitkin. (2015), 'Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation', Journal of Computational and Graphical Statistics 24, 44–65.

Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy. (2014), 'Explaining and Harnessing Adversarial Examples', arXiv Preprint, 1412.6572.

Government of Canada (2020), 'Algorithmic Impact Assessment', https://canada-ca.github.io/aia-eia-js, accessed 24 April 2022.

Greenwell, Brandon M., Bradley C. Boehmke, and Andrew J. McCarthy. (2018), 'A Simple and Effective Model-Based Variable Importance Measure', arXiv Preprint, 1805.04755.

Gregor, Shirley, and Izak Benbasat. (1999), 'Explanations from Intelligent Systems: Theoretical Foundations and Implications for Practice', MIS Quarterly 23, 497–530.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. (2018), 'A Survey of Methods for Explaining Black Box Models', ACM Computing Surveys 51, 93:1–93:42. doi: https://doi.org/10.1145/3236009.

Hempel, Carl G. (1965), Aspects of Scientific Explanation. (London: Collier-MacMillan Limited).

Hempel, Carl G., and Paul Oppenheim. (1948), 'Studies in the Logic of Explanation', Philosophy of Science 15, 135–175.

Henelius, Andreas, Kai Puolamäki, Henrik Boström, Lars Asker, and Panagiotis Papapetrou. (2014), 'A Peek into the Black Box: Exploring Classifiers by Randomization', Data Mining and Knowledge Discovery 28, 1503–1529. doi: https://doi.org/10.1007/s10618-014-0368-8.

Hesse, Mary B. (1965), Models and Analogies in Science (London: Sheed and Ward).

Hesslow, Germund. (1988), 'The Problem of Causal Selection', in D.J. Hilton, ed, Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality (Brighton: Harvester Press), 11–31.

High Level Expert Group on Artificial Intelligence (2019), Ethics Guidelines for Trustworthy AI (European Commission).

Hilton, Denis J. (1990), 'Conversational Processes and Causal Explanation', Psychological Bulletin 107, 65.

Hilton, Denis J. (1996), 'Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance', Thinking & Reasoning 2, 273–308.

Hilton, Denis J., and Ben R. Slugoski. (1986), 'Knowledge-Based Causal Attribution: The Abnormal Conditions Focus Model', Psychological Review 93, 75.

Holland, Sarah, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. (2018), 'The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards', arXiv Preprint, 1805.03677 [cs].

Hooker, Giles. (2004), 'Discovering Additive Structure in Black Box Functions', in Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge

Discovery and Data Mining (New York: Association for Computing Machinery), 575–580.

Information Commissioner's Office, The Alan Turing Institute (2020), Explaining Decisions Made with AI.

Jobin, Anna, Marcello Ienca, and Effy Vayena. (2019), 'The Global Landscape of AI Ethics Guidelines', Nature Machine Intelligence 1, 389–399. doi: https://doi.org/10.1038/s42256-019-0088-2.

Kayande, Ujwal, Arnaud De Bruyn, Gary L. Lilien, Arvind Rangaswamy, and Gerrit H. Van Bruggen. (2009), 'How Incorporating Feedback Mechanisms in a DSS Affects DSS Evaluations', Information Systems Research 20, 527–546.

Kim, Been, Cynthia Rudin, and Julie A. Shah. (2014), 'The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification', in Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (Montreal: MIT Press), 1952–1960.

Kim, Been, Rajiv Khanna, and Oluwasanmi O. Koyejo. (2016), 'Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability', Advances in Neural Information Processing Systems, 2288–2296.

Kment, Boris. (2006), 'Counterfactuals and Explanation', Mind 115, 261–310.

Koh, Pang Wei, and Percy Liang. (2017), 'Understanding Black-Box Predictions via Influence Functions', arXiv Preprint, 1703.04730 [cs, stat].

Krause, Josua, Adam Perer, and Kenney Ng. (2016), 'Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models', in Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16 (New York: Association for Computing Machinery), 5686–5697. doi: https://doi.org/10.1145/2858036.2858529.

Krishnan, R., G. Sivakumar, and P. Bhattacharya. (1999), 'Extracting Decision Trees from Trained Neural Networks', Pattern Recognition 32, 1999–2009. doi: https://doi.org/10.1016/S0031-3203(98)00181-2.

Kulesza, Todd, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. (2015), Principles of Explanatory Debugging to Personalize Interactive Machine Learning (New York: Association for Computing Machinery Press). doi: https://doi.org/10.1145/2678025.2701399.

Lakkaraju, Himabindu, and Osbert Bastani. (2019), '"How Do I Fool You?": Manipulating User Trust via Misleading Black Box Explanations', arXiv Preprint, 1911.06473.

Lakkaraju, Himabindu, Stephen H. Bach, and Jure Leskovec. (2016), 'Interpretable Decision Sets: A Joint Framework for Description and Prediction', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 (New York: Association for Computing Machinery), 1675–1684. doi: https://doi.org/10.1145/2939672.2939874.

Lakkaraju, Himabindu, Ece Kamar, Rich Caruana, and Jure Leskovec. (2017), 'Interpretable & Explorable Approximations of Black Box Models', arXiv Preprint, 1707.01154 [cs].

Lepri, Bruno, Nuria Oliver, Emmanuel Letouzé, Alex Pentland, and Patrick Vinck. (2017), 'Fair, Transparent, and Accountable Algorithmic Decision-Making Processes: The Premise, the Proposed Solutions, and the Open Challenges', Philosophy & Technology 31. doi: https://doi.org/10.1007/s13347-017-0279-x.

Lewis, David. (1973), Counterfactuals (Oxford: Blackwell).

Lewis, David. (1986), Philosophical Papers II (Oxford: Oxford University Press).

Lim, Brian Y., and Anind K. Dey. (2009), 'Assessing Demand for Intelligibility in Context-Aware Applications', in Proceedings of the 11th International Conference on

Ubiquitous Computing - Ubicomp ’09, presented at the the 11th international conference (Orlando, FL: ACM Press), 195. doi: https://doi.org/10.1145/1620545.1620576.

Lipton, Peter. (1990), ‘Contrastive Explanation’, Royal Institute of Philosophy Supplements 27, 247–266.

Lipton, Peter. (2001), ‘What Good is an Explanation?’, in Explanation (Springer: Dordrecht), 43–59.

Lipton, Zachary C. (2016), ‘The Mythos of Model Interpretability’, arXiv Preprint, 1606.03490 [cs, stat].

Lisboa, Paulo JG. (2013), ‘Interpretability in Machine Learning—Principles and Practice’, in Fuzzy Logic and Applications (Springer: Cham), 15–21.

Lombrozo, Tania. (2009), ‘Explanation and Categorization: How “Why?” Informs “What?”’, Cognition 110, 248–253.

Lou, Yin, Rich Caruana, and Johannes Gehrke. (2012), ‘Intelligible Models for Classification and Regression’, in Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’12 (New York: Association for Computing Machinery), 150–158. doi: https://doi.org/10.1145/2339530.2339556.

Lou, Yin, Rich Caruana, Johannes Gehrke, and Giles Hooker. (2013), ‘Accurate Intelligible Models with Pairwise Interactions’, in Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’13 (New York: Association for Computing Machinery), 623–631. doi: https://doi.org/10.1145/2487575.2487579.

Lundberg, Scott, and Su-In Lee. (2017), ‘A Unified Approach to Interpreting Model Predictions’, arXiv Preprint, 1705.07874 [cs, stat].

Manders-Huits, Noëmi, and Michael Zimmer. (2009), 'Values and Pragmatic Action: The Challenges of Introducing Ethical Intelligence in Technical Design Communities', International Review of Information Ethics 10, 37–44.

Mantelero, Alessandro. (2016), 'Personal Data for Decisional Purposes in the Age of Analytics: From an Individual to a Collective Dimension of Data Protection', Computer Law & Security Review 32, 238–255. doi: https://doi.org/10.1016/j.clsr.2016.01.014.

Mantelero, Alessandro. (2018), 'AI and Big Data: A Blueprint for a Human Rights, Social and Ethical Impact Assessment', Computer Law & Security Review 34, 754–772.

Martens, David, and Foster Provost. (2014), 'Explaining Data-Driven Document Classifications', MIS Quarterly 38, 73-100.

McAuley, Julian, and Jure Leskovec. (2013), 'Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text', Proceedings of the 7th ACM Conference on Recommender Systems (ACM Press: New York), 165–172. doi: https://doi.org/10.1145/2507157.2507163.

McClure, John L., Robbie M. Sutton, and Denis J. Hilton. (2003), 'Implicit and explicit processes in social judgements: The role of goal-based explanations', in Joseph P. Forgas, Kipling D. Williams, and William von Hippel, eds., Social Judgements: Implicit and Explicit Processes (Cambridge: Cambridge University Press), 306-324.

Mercier, Hugo, and Dan Sperber. (2011), 'Why Do Humans Reason? Arguments for an Argumentative Theory', Behavioral and Brain Sciences 34, 57-74.

Miller, Tim. (2019), 'Explanation in Artificial Intelligence: Insights from the Social Sciences', Artificial Intelligence 267, 1–38. doi: https://doi.org/10.1016/j.artint.2018.07.007.

Mitchell, Margaret, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. (2019), 'Model Cards for Model Reporting', Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, 220–229. doi: https://doi.org/10.1145/3287560.3287596.

Mittelstadt, Brent. (2016), 'Auditing for Transparency in Content Personalization Systems', International Journal of Communication 10, 12.

Mittelstadt, Brent. (2017), 'From Individual to Group Privacy in Big Data Analytics', Philosophy & Technology 30, 475–494. doi: https://doi.org/10.1007/s13347-017-0253-7.

Mittelstadt, Brent. (2019), 'Principles Alone Cannot Guarantee Ethical AI', Nature Machine Intelligence 1, 501–507. doi: https://doi.org/10.1038/s42256-019-0114-4.

Mittelstadt, Brent, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. (2016), 'The Ethics of Algorithms: Mapping the Debate', Big Data & Society 3. doi: https://doi.org/10.1177/2053951716679679.

Mittelstadt, Brent, Chris Russell, and Sandra Wachter. (2019), 'Explaining Explanations in AI', Proceedings of the Conference on Fairness, Accountability, and Transparency—FAT* '19, 279–288. doi: https://doi.org/10.1145/3287560.3287574.

Molnar, Christoph. (2020), Interpretable Machine Learning.

Montavon, Grégoire, Wojciech Samek, and Klaus-Robert Müller. (2018), 'Methods for Interpreting and Understanding Deep Neural Networks', Digital Signal Processing 73, 1-15.

Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. (2016), 'Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks', in Daniel D. Lee, Masashi Sugiyama, Ulrike Von Luxburg,

Isabelle Guyon, and Roman Garnett, eds, Advances in Neural Information Processing Systems 29 (Curran Associates, Inc: New York), 3387–3395.

Pearl, Judea. (2000), Causation (Cambridge: Cambridge University Press).

Pearl, Judea. (2019), 'The Seven Tools of Causal Inference, with Reflections on Machine Learning', Communications of the ACM 62, 54–60.

Põder, T., and T. Lukki. (2011), 'A Critical Review of Checklist-Based Evaluation of Environmental Impact Statements', Impact Assessment and Project Appraisal 29, 27–36.

Poulin, Brett, Roman Eisner, Duane Szafron, Paul Lu, Russ Greiner, D. S. Wishart, Alona Fyshe, Brandon Pearcy, Cam MacDonell, and John Anvik. (2006), 'Visual Explanation of Evidence in Additive Classifiers', in Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence, Vol. 2, IAAI '06 (Boston, MA: AAAI Press), 1822–1829.

Putnam, Hilary. (1978), Meaning and the Moral Sciences (London: Routledge & Kegan Paul).

Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. (2018), Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. (2016), '"Why Should I Trust You?": Explaining the Predictions of Any Classifier', in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16 New York, NY: ACM Press), 1135–1144. doi: https://doi.org/10.1145/2939672.2939778.

Ruben, David-Hillel. (2004), Explaining Explanation (Routledge: New York).

Rudin, Cynthia. (2019a), 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', arXiv Preprint, 1811.10154 [cs, stat].

Rudin, Cynthia. (2019b), 'Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead', Nature Machine Intelligence 1, 206–215. doi: https://doi.org/10.1038/s42256-019-0048-x.

Russell, Chris. (2019), 'Efficient Search for Diverse Coherent Explanations', in Proceedings of the Conference on Fairness, Accountability, and Transparency (New York: ACM Press), 20–28.

Salmon, Wesley C. (2006), Four Decades of Scientific Explanation (University of Pittsburgh Press: Pittsburgh).

Saltelli, Andrea. (2002), 'Sensitivity Analysis for Importance Assessment', Risk Analysis 22, 579–590. doi: https://doi.org/10.1111/0272-4332.00040.

Samek, Wojciech, Thomas Wiegand, and Klaus-Robert Müller. (2017), 'Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models', arXiv Preprint, 1708.08296.

Samland, Jana, and Michael R. Waldmann. (2014), 'Do Social Norms Influence Causal Inferences?', Proceedings of the Annual Meeting of the Cognitive Science Society 36, 1359-1364.

Sandvig, Christian, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. (2014), 'Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms', Data and Discrimination: Converting Critical Concerns into Productive Inquiry.

Schölkopf, Bernhard. (2019), 'Causality for Machine Learning', arXiv Preprint, 1911.10500.

Selvaraju, Ramprasaath R., Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. (2016), 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization', v3, https://doi.org/10.48550/arXiv.1610.02391, accessed 24 April 2022.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. (2013), 'Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps', arXiv Preprint, 1312.6034.

Slugoski, Ben R., Mansur Lalljee, Roger Lamb, and Gerald P. Ginsburg. (1993), 'Attribution in Conversational Context: Effect of Mutual Knowledge on Explanation-Giving', European Journal of Social Psychology 23, 219–238.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. (2013), 'Intriguing Properties of Neural Networks', arXiv Preprint, 1312.6199.

Tamagnini, Paolo, Josua Krause, Aritra Dasgupta, and Enrico Bertini. (2017), Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations (ACM Press: New York). doi: https://doi.org/10.1145/3077257.3077260.

Thiagarajan, Jayaraman J., Bhavya Kailkhura, Prasanna Sattigeri, and Karthikeyan Natesan Ramamurthy. (2016), 'TreeView: Peeking into Deep Neural Networks Via Feature-Space Partitioning', arXiv Preprint, 1611.07429 [cs, stat].

Turner, Ryan. (2016), 'A Model Explanation System', in 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), presented at the 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 1–6. doi: https://doi.org/10.1109/MLSP.2016.7738872.

Wachter, Sandra, and Brent Mittelstadt. (2019), ‘A Right to Reasonable Inferences: Re-Thinking Data Protection Law in the Age of Big Data and AI’, Columbia Business Law Review 2, 494-620.

Wachter, Sandra, Brent Mittelstadt, and Luciano Floridi. (2017), ‘Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation’, International Data Privacy Law 7, 76–99.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. (2018), ‘Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR’, Harvard Journal of Law & Technology 3, 841–887.

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. (2020), ‘Why Fairness Cannot Be Automated: Bridging the Gap between EU Non-Discrimination Law and AI’, SSRN Scholarly Paper No. ID 3547922) (Rochester, NY: Social Science Research Network). doi: https://doi.org/10.2139/ssrn.3547922.

Wang, Fulton, and Cynthia Rudin. (2015), ‘Falling Rule Lists’, in Lebanon, Guy and Vishwanathan, S. V. N., eds., Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics (San Diego: PMLR), 1013–1022.

Wexler, James, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. (2019), ‘The What-If Tool: Interactive Probing of Machine Learning Models’, IEEE Transactions on Visualization and Computer Graphics 1, 1. https://doi.org/10.1109/TVCG.2019.2934619

Woodward, James. (2005), Making Things Happen: A Theory of Causal Explanation (Oxford: Oxford University Press).

Woodward, James. (2003), ‘Scientific Explanation’, in Edward N. Zalta, ed., The Stanford Encyclopedia of Philosophy (Stanford: Stanford University). URL: https://plato.stanford.edu/archives/sum2003/entries/scientific-explanation/

Woodward, James. (1997), ‘Explanation, invariance, and intervention’, Philosophy of Science 64.S4, S26–S41.

Yang, Ke, Julia Stoyanovich, Abolfazl Asudeh, Bill Howe, Hv Jagadish, and Gerome Miklau. (2018), ‘A Nutritional Label for Rankings’, in Proceedings of the 2018 International Conference on Management of Data - SIGMOD ’18, presented at the the 2018 International Conference (Houston, TX: ACM Press), 1773–1776. doi: https://doi.org/10.1145/3183713.3193568.

Yin, Xiaoxin, and Jiawei Han. (2003), ‘CPAR: Classification Based on Predictive Association Rules’, in Proceedings of the 2003 SIAM International Conference on Data Mining, Proceedings (Society for Industrial and Applied Mathematics), 331–335. doi: https://doi.org/10.1137/1.9781611972733.40.

Ylikoski, Petri. (2013) ‘Causal and Constitutive Explanation Compared’, Erkenntnis 78, 277–297.

Yosinski, Jason, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. (2015), ‘Understanding Neural Networks through Deep Visualization’, arXiv Preprint, 1506.06579 [cs].

Zarsky, Tal Z. (2013), ‘Transparent Predictions’, Univeristy of Illinois Law Review 4, 1503.

[1] The degree to which the reasons for specific model behaviours can be explained is sometimes referred to as the *explainability* of a model. Here, it is treated as one component of *interpretability* alongside intrinsic model comprehensibility.

[2] Lessons can be learned from comparable legal instruments. For example, ‘data protection impact assessments’ as required by Article 35 of the EU General Data Protection Regulation are functionally similar to the self-assessment frameworks discussed above but do not require full public disclosure of results.