Explainability Is in the Mind of the Beholder: Establishing the Foundations of Explainable Artificial Intelligence

KACPER SOKOL, ARC Centre of Excellence for Automated Decision-Making and Society, School of Computing Technologies, RMIT University, Australia and Department of Computer Science, University of Bristol, United Kingdom
PETER FLACH, Department of Computer Science, University of Bristol, United Kingdom

Explainable artificial intelligence and interpretable machine learning are research fields growing in importance. Yet, the underlying concepts remain somewhat elusive and lack generally agreed definitions. While recent inspiration from social sciences has refocused the work on needs and expectations of human recipients, the field still misses a concrete conceptualisation. We take steps towards addressing this challenge by reviewing the philosophical and social foundations of human explainability, which we then translate into the technological realm. In particular, we scrutinise the notion of algorithmic black boxes and the spectrum of understanding determined by explanatory processes and explainees' background knowledge. This approach allows us to define explainability as (logical) *reasoning* applied to transparent *insights* (into black boxes) interpreted under certain *background knowledge* – a process that engenders *understanding* in explainees. We then employ this conceptualisation to revisit the much disputed trade-off between transparency and predictive power and its implications for ante-hoc and post-hoc explainers as well as fairness and accountability engendered by explainability. We furthermore discuss components of the machine learning workflow that may be in need of interpretability, building on a range of ideas from human-centred explainability, with a focus on explainees, contrastive statements and explanatory processes. Our discussion reconciles and complements current research to help better navigate open questions – rather than attempting to address any individual issue – thus laying a solid foundation for a grounded discussion and future progress of explainable artificial intelligence and interpretable machine learning. We conclude with a summary of our findings, revisiting the human-centred explanatory process needed to achieve the desired level of algorithmic transparency.

Additional Key Words and Phrases: Defining Explainability, Interpreting and Understanding Explanations, Human-Centred Perspective.

1 GREAT EXPECTATIONS: EXPLAINABLE MACHINES

Transparency, interpretability and explainability promote understanding and confidence. As a society, we strive for transparent governance and justified actions that can be scrutinised and contested. Such a strong foundation provides a principled mechanism for reasoning about fairness and accountability, which we have come to expect in many areas. However, Artificial Intelligence (AI) systems are not always held to the same standards. This becomes problematic when such systems power applications that either implicitly or explicitly affect people's lives, for example in banking, justice, job screenings or school admissions [69]. In such cases, creating explainable predictive models or retrofitting transparency to pre-existing algorithms is usually expected by the affected individuals, or simply enforced by law. A number of techniques and algorithms are being proposed to this end; however, as a relatively young research area, there is no consensus within the AI discipline on a suite of technology to address these challenges.

Building intelligible AI systems is not unproblematic, particularly given their varied, and sometimes ambiguous, audience [42, 75], purpose [34] and application domain [9]. While intelligent systems are often deemed (unconditionally) opaque, it is not a definitive property and it largely depends on all of the aforementioned factors, some of which fall beyond the consideration of a standard AI development lifecycle. Without clearly defined explainability desiderata [89] addressing such aspects can be challenging, in contrast to designing AI systems purely based on their predictive performance, which is often treated as a quality proxy that can be universally measured, reported and compared. In

2021. Manuscript submitted to ACM

view of this disparity, many engineers (incorrectly) consider these two objectives as competing [79], thus choosing to pursue high predictive performance at the expense of opaqueness, which may be incentivised by business opportunities.

While high predictive power of an AI system might make it useful, its explainability determines its acceptability. The pervasiveness of automated decision-making in our everyday life, some of them bearing social consequences, requires striking a balance between the two that is appropriate for what is at stake; for example, approaching differently a car autopilot and an automated food recommendation algorithm. Another domain that could benefit from powerful and explainable AI is (scientific) discovery – intelligent systems may achieve super-human performance, e.g., AlphaGo [84], however a lack of transparency renders their mastery and ingenuity unattainable. Such observations have prompted the Defense Advanced Research Projects Agency (DARPA) to announce the eXplainable AI (XAI) programme [32, 33] that promotes building a suite of techniques to (i) create more explainable models while preserving their high predictive performance; and (ii) enable humans to understand, trust and effectively manage intelligent systems.

To address these challenges, AI explainability and Machine Learning (ML) interpretability solutions are developed at breakneck speed, giving a perception of a chaotic field that may seem difficult to navigate at times. Despite these considerable efforts, a universally agreed terminology and evaluation criteria are still elusive, with many methods introduced to solve a commonly acknowledged but under-specified problem, and their success judged based on ad-hoc measures. In this paper we take a step back and re-evaluate the foundations of this research to organise and reinforce its most prominent findings that are essential for advancing these fields, with the aim of providing a well-grounded platform for tackling DARPA's XAI mission. Our work thus reconciles and complements the already vast body of technical contributions, philosophical and social treatises, as well as literature surveys by bringing to light the interdependence of interdisciplinary and multifaceted concepts upon which explainable AI and interpretable ML are built. Our discussion manoeuvres through this incoherent landscape by connecting numerous open questions and challenges – rather than attempting to "solve" any individual issue – which is achieved through a comprehensive review of published work that acknowledges any difficulties or disagreements pertaining to the field.

In particular, we review the notions of *black box* and *opaqueness* in the context of artificial intelligence, and formalise *explainability* – the preferred term in our approach (Section 2). We discuss the meaning and purpose of explanations, and identify conceptual prerequisites for lifting unintelligibility of predictive systems, based on which we define explainability as a technology providing insights that lead to understanding, which both conceptualises such techniques and fixes their evaluation criterion. Furthermore, we show how transparency, and other terms often used to denote similar concepts, can be differentiated from explainability – both overcome opaqueness, but only the latter leads to understanding – which we exemplify with decision trees of different sizes. While the premise of our definition is clear, understanding largely depends upon the explanation recipients, who come with a diverse range of background knowledge, mental models, expectations and experience. Therefore, in addition to technical requirements, explainability tools should also embody various social traits as their output is predominantly aimed at humans. We discuss these aspects of AI and ML explainers in Section 3, which considers the role of an *explanation audience* and their preference for *contrastive statements* – XAI insights inspired by explainability research in the social sciences [64]. We also examine the social and bi-directional *explanatory process* underlying conversational explanations between humans, highlighting the desire of an explainee to customise, personalise and contest various aspects of explanations provided for opaque systems within a congruent interaction [91, 95].

Next, in Section 4 we take a closer look at explainability by design (ante-hoc) and techniques devised to remove opaqueness from pre-existing black boxes (post-hoc and, often, model-agnostic), focusing on the latter type given that such methods are universally applicable to a wide variety models, which increases their potential reach and impact.

While these explainers are appealing, their modus operandi can be an unintended cause of low-fidelity explanations that lack truthfulness with respect to the underlying black box [79]. Furthermore, their flexibility means that, from a technical perspective, they can be applied to any predictive model, however they may not necessarily be equally well suited to the intricacies of each and every one of them. Creating a faithful post-hoc explainer requires navigating multiple trade-offs reflected in choosing specific components of otherwise highly-modular explainability framework and parameterising these building blocks based on the specific use case [90, 92, 96]. These observations prompt us to revisit the disputed *transparency–predictive power trade-off* and assess *benefits* of interpretability that go beyond understanding of predictive algorithms and their decisions, such as their *fairness* and *accountability*.

We continue our investigation in Section 5 by assessing *explainability needs of various parts of predictive systems* – data, models and predictions – as well as multiplicity and diversity of these, sometimes incompatible, insights. To this end, we use an *explainability taxonomy* derived from the Explainability Fact Sheets [89] to reason about such systems within a well-defined framework that considers both their social and technical requirements. Notably, it covers human aspects of explanations, thus giving us a platform to examine the audience (explainees), explanation complexity and fidelity, as well as the interaction mode, among many others. This discussion leads us to a high-level overview of landmark explainable AI and Interpretable ML (IML) literature that highlights the interplay between various (often interdisciplinary and multifaceted) concepts popular in the field, thus painting a coherent perspective. Section 6 then summarises our main observations and contributions:

- we formally define explainability and catalogue other relevant nomenclature;
- we establish a spectrum of "opaqueness" determined by the desired level of transparency and interpretability;
- we identify important gaps in human-centred explainability from the perspective of current technology;
- we dispute universality of post-hoc explainers given their complexity and high degree of modularity; and
- we address explanation multiplicity through explanatory protocols for data, models and predictions.

These insights pave the way for the development of more intelligible and robust machine learning and artificial intelligence explainers.

2 DEFINING BLACK-BOX AND EXPLAINABLE ARTIFICIAL INTELLIGENCE

To avoid a common criticism of explainability research, we begin by discussing the concept of interpretability. To this end, we identify causes of opaqueness when dealing with intelligent systems and assess prerequisites for their understanding. In this setting we observe **shades of black-boxiness**: an interpretability spectrum determined by the extent of understanding exhibited by explainees, which, among others, is conditioned upon their mental capacity and background knowledge. We link this finding with various notions used in XAI and IML literature, a connection that helps us to fix the nomenclature and **define explainability** (our preferred term).

The term **black box** can be used to describe a system whose internal workings are opaque to the observer – its operation may only be traced by analysing its inputs and outputs [7, 13]. Similarly, in computer science (including AI and ML) a black box is a (data-driven) algorithm that can be understood as an automated process that we cannot reason about beyond observing its behaviour. For AI in particular, Rudin [79] points out two main sources of opaqueness: (i) a *proprietary* system, which may be transparent to its creators, but operates as a black box; and (ii) a system that is too *complex* to be comprehend by *any human*. While the latter case concerns systems that are universally opaque for the *entire population*, we argue that – in contrast to a binary classification [18] – this definition of black boxes essentially establishes a *spectrum of understanding*.

Similarly, a seminal inquiry into opaqueness of visual perception systems by Marr [60] suggested three different levels at which information processing devices can be understood. The top tier is *computational theory*, which concerns abstract specification of the problem at hand and the overall goal. It is followed by *representation and algorithm*, which deals with implementation details and selection of an appropriate representation. The final level is *hardware implementation*, which simply establishes physical realisation of the explained problem. To illustrate his framework, Marr [60] argued that understanding why birds fly cannot be achieved by only studying their feathers: "In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense." Nonetheless, he points out that these three levels are only loosely related and some phenomena may be explained at only one or two of them, therefore it is important to identify which of these levels need to be covered in each individual case to arrive at understanding.

Undeniably, perception and comprehension of a phenomenon depend upon the observer's cognitive capabilities and mental model, the latter of which is an internal representation of this phenomenon built on real-world experiences [48]. For example, Kulesza et al. [48] outline a *fidelity*-based understanding spectrum spanning two dimensions:

completeness how truthful the understanding is overall (generality); and
soundness how accurate the understanding is for a particular phenomenon (specificity).

Therefore, a *complete* understanding of an event from a certain domain is equivalently applicable to other, possibly unrelated, events from the same domain; for example, gravity in relation to a pencil falling from a desk. A *sound* understanding, on the other hand, accurately describes an event without (over-)simplifications, which may result in misconceptions; for example, leaving a pencil on a slanted surface results in it falling to the ground. Striking the right balance between the two depends upon the observer and may be challenging to achieve: completeness without soundness is likely to be too broad, hence uninformative; and the opposite can be too specific to the same effect.

Within this space, Kulesza et al. [48] identify two particularly appealing types of a mental model:

functional which is enough to operationalise a concept but does not necessarily entail the understanding of its underlying mechanism (akin to The Chinese Room Argument [72, 83]); and
structural which warrants a detailed understanding of how and why a concept operates.

For example, a functional understanding of a switch and a light bulb circuit can be captured by the dependency between flipping the switch and the bulb lighting up. A structural understanding of the same phenomenon, on the other hand, may focus on the underlying physical processes, e.g., closing an electrical circuit allows electrons to move, which heats up the bulb's filament, thus emitting light. The former understanding is confined to operating a light switch, while the latter can be generalised to many other electrical circuits. Each one is aimed at a different audience and their complexity should be fine-tuned for the intended purpose as explanations misdirected towards an inappropriate audience may be incomprehensible. We argue that this spectrum of understanding in human explainability can constitute a yardstick for determining explanatory qualities of predictive algorithms – a link that has mostly been neglected in the literature, but which can help us to explicitly define popular XAI and IML terminology.

A very considerable amount of research into explainable AI and interpretable ML published in recent years may suggest that it is a freshly established field, however in reality it is more of a renaissance. While work in this area indeed picked up the pace in the past decade, interest in creating transparent and explainable, data-driven algorithms dates back at least to the 1990s [79], and further back to the 1970s if expert systems [53] are taken into account. With such a rich history and the increased publication velocity attributed to the more recent re-establishment of the field, one may think that this research area has clearly defined objectives and a widely shared and adopted **terminology**. However,

with an abundance of keywords that are often used interchangeably in the literature – without precisely defining their meaning – this is not yet the case. The most common terms include, but are not limited to:

- explainability,
- intelligibility,
- justification,
- observability,
- comprehensibility,
- evidence,
- transparency,
- interpretability,
- reason, and
- explicability,
- simulatability,
- cause.

While early research might have missed out on an opportunity to clearly define its goals and nomenclature, recent work [3, 10, 11, 27, 66, 79] has attempted to rectify this problem. Gilpin et al. [27] offered definitions of "explanation", "interpretability" and "explainability" drawing from a broad body of literature in an effort to standardise XAI and IML work. While their notions appear somewhat vague – explanations should answer "Why?" and "Why-should?" questions until such questions can no longer be asked – they argue for making explanations *interpretable* and *complete*, striving towards human *understanding* that depends on the explainee's cognition, knowledge and biases. Similarly, Biran and McKeown [11] were concerned with *explanations*, which they characterised as "giving a reason for a prediction" and answering "how a system arrives at its prediction". They also defined *justifications* as "putting an explanation in a context" and conveying "why we should believe that the prediction is correct", which, they note, do not necessarily have to correspond to how the predictive system actually works. In a later piece of work, Biran and Cotton [10] defined explainability as "the degree to which an observer can understand the cause of a decision" (also adopted by Miller [64]), thus making it much more explainee-centred. Another important term is *cause*; while it is used sparingly in the XAI and IML literature, it should be reserved exclusively for insights extracted from causal models [71].

More recently, based on an extensive review of literature in computer science and related disciplines, Mohseni et al. [66] provided a collection of definitions for the most common terms in XAI and IML, nonetheless the underlying rationale is predominantly qualitative making them difficult to operationalise. In contrast, Alvarez-Melis et al. [3] used the *weight of evidence* concept from information theory to mathematically define AI and ML explainability and outline a precise list of its desiderata. While appealing, the complexities of the real world make their framework difficult to apply at large. In another attempt, Rudin [79] defined *interpretability* as a domain-specific notion that imposes "a set of application-specific constraints on the model", thus making this concept only applicable to predictive models that can provide their own explanations (i.e., ante-hoc interpretability). Therefore, in her view a predictive model is interpretable if it "obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity or physical constraints that come from domain knowledge". Rudin also objects to using the term *explanation* when referring to "approximations to black box model predictions" (i.e., post-hoc explainability).

Each definition conveys a more or less precise meaning that can be used to label relevant techniques, however they do not necessarily clarify and help to navigate the complex landscape of IML and XAI research. To organise this space, we categorise the underlying terminology based on three criteria:

- *properties* of systems,
- *actions* required to process and assimilate them.
- *functions* and *roles* which they serve, and

The core concept around which we build our nomenclature is **explainability**; we define it as **insights that lead to understanding** (the **role** of an explanation) – a popular rationale in the social sciences [6, 43, 56]. While it may seem abstract, understanding can be assessed with questioning dialogues [5, 59, 100–102] – e.g., a machine interrogating the explainees to verify their understanding of the phenomenon being explained – which are the opposite of explanatory dialogues. Such a process reflects how understanding is tested in education, where the quality of tuition and knowledge of pupils is evaluated through standardised tests and exams (albeit not without criticism [62]). Furthermore, encouraging

people to explain a phenomenon helps them to realise the extent of their ignorance and confront the complexity of the problem, which are important factors in exposing The Illusion of Explanatory Depth [78] – a belief that one understands more than one actually does.

This notion of explainability and the three building blocks of XAI and IML terminology allow us to precisely define the other popular terms. Therefore,

- *observability*,
- *transparency*,
- *explicability*,
- *intelligibility*,
- *comprehensibility*, and
- *interpretability*

are **properties** of an AI system that enable it to convey information of varied complexity, the *understanding* of which depends upon the cognitive capabilities and (domain) expertise of the explainee. For example, observing an object falling from a table is a transparent phenomenon per se, but the level of its understanding, if any, is based upon the depth of the observer's physical knowledge. Such transparency provides

- *evidence*,
- *reason*, and
- *justification*

(**roles**) that can be used to

- *reason* about,
- *interpret*, or
- *comprehend*

(note that here the three are used as verbs) behaviour of a black box, all three of which are **actions** that possibly lead to understanding. While *simulatability* is also based upon observing a transparent behaviour and replicating it, such an **action** does not necessarily imply understanding of the underlying phenomenon – recall the difference between structural and functional mental models [48] and The Chinese Room Argument [83] discussed earlier. Lastly, a *cause* has a similar meaning to a *reason*, but the first one is derived from a proper causal model, whereas the latter is based purely on observations of the behaviour of a black-box model.

Such a setting makes a welcome connection between the XAI and IML terminology summarised by the equation

$$\texttt{Explainability} = \underbrace{\texttt{Reasoning}\,(\texttt{Transparency} \mid \texttt{Background Knowledge})}_{\textit{understanding}},$$

which defines `Explainability` as the **process** of deriving *understanding* through `Reasoning` applied to `Transparent` insights extracted from the black box that are adjusted to the explainee's `Background Knowledge`. In this process, the `Reasoning` can either be done by the explainer or the explainee, and there is an implicit assumption that the explainee's `Background Knowledge` aligns with the `Transparent` representation of the black box. If the latter does not hold, mitigation techniques such as employing an *interpretable representation* can be used to communicate concepts that are otherwise incomprehensible [76, 92, 96]. `Reasoning` also comes in many different shapes and sizes depending on the underlying black box (`Transparency`) as well as the explainer and the explainee (`Background Knowledge`); for example, logical reasoning with facts, causal reasoning over a causal graph, case-based reasoning with a fixed similarity metric, and artificial neuron activation analysis for a *shallow* neural network.

Therefore, linear models are transparent given a reasonable number of features; additionally, with the right ML and domain background knowledge – requirement of normalised features, effect of feature correlation and the meaning of coefficients – the explainee can reason about their properties, leading to an explanation based on understanding. Similarly, a visualisation of a *shallow* decision tree can be considered both transparent and explainable assuming that the explainee understands how to navigate its structure (ML background knowledge) and the features are meaningful (domain background knowledge); again, it is up to the explainee to reason about these insights. When the size of a tree

increases, however, its visualisation loses the explanatory power because many explainees become unable to process and reason about its structure. Restoring the explainability of a deep tree requires delegating the reasoning process to an algorithm that can digest its structure and output sought after insights in a concise representation. For example, when explaining a prediction, the tree structure can be traversed to identify a similar instance with a different prediction, e.g., as encoded by two neighbouring leaves with a shared parent, thus demystifying the automated decision [88].

3 HUMANS AND EXPLANATIONS: TWO SIDES OF THE SAME COIN

Defining explainability as "leading to understanding" and our categorisation into *properties*, *functions* and *actions* highlight an important aspect of this research topic: explanations are directed at some autonomous agent, either a human or machine, who is as important as the explainability algorithm itself. Notably, up until recently XAI and IML research has been undertaken mostly within the computer science realm [65], thus bringing in various biases and implicit assumptions from this predominantly technical field. While some explainability research has found its way into other scientific disciplines, e.g., law [99], the majority gravitated around technical properties. This research agenda was disrupted by Miller et al. [65], who observed that the function of an explanation and its recipients are largely neglected – a phenomenon which they dubbed "inmates running the asylum" – leading to a substantial paradigm shift. Miller's [64] follow-on work grounded this observation in (human) explainability research from the social sciences, where this topic has been studied for decades, thus providing invaluable insights that can benefit XAI and IML.

Miller's findings [64] have arguably reshaped the field, with a substantial proportion of the ensuing research acknowledging the **explainees** – their goals, expectations, intentions and interactions. While explainability of autonomous systems has various benefits, it is usually requested when an AI agent operates inconsistently with the explainee's expectations or mental model, e.g., an unexpected ML prediction causing a disagreement. In such a case, explainees' preferences, needs and goals should be addressed to maximise the effectiveness of an explanation, for example by appropriately adjusting its complexity [64]. This step can be further improved by treating explainability as a process instead of one-off information offloading [64]; by satisfying the explainees' natural desire to **interact** and communicate with the explainer within a predictable protocol, they are provided with an opportunity to customise and personalise the explanation [91]. Perhaps the most influential of Miller's observations is the humans' preference for **contrastive** explanations given their prominence in our everyday life. We discuss these three fundamental aspects of human-centred explainability in more detail below.

Understanding can be an elusive objective when it comes to explaining intelligent systems since each unique **explanation audience** may expect to receive different insights, e.g., a medical diagnosis can be communicated in terms of test results or observable symptoms depending on whether it is directed towards medical staff or patients. While in our considerations we implicitly assume that the explanation recipient is a human, it may as well be another algorithm that further processes such insights, in which case other, more suitable, properties would be of interest. When taken into account, the *purpose* of explainability and the explainee's goal also influence the explanation composition. For example, an explanation will look different when it helps to debug an ML model and when it justifies a negative outcome of a loan application; note that the target audience also differs, with the former aimed at ML engineers and the latter at lay people. Addressing such desiderata by accounting for the explainee's cognitive capabilities and skill level, however, is challenging as it requires access to the explainee's background knowledge and mental model, which are vague and often undefined concepts that cannot be easily extracted and modelled.

Nonetheless, just by considering the audience and purpose of an explanation, we can identify (and deliver) a collection of relevant properties. In certain cases, such as the aforementioned loan application, the *actionability* of explanatory

insights is crucial, e.g., suggesting that an individual would receive a loan had he or she been 10 years younger is futile. Multiplicity of apparently indistinguishable arguments can also decrease the perceived quality of an explanation when one is chosen at random without a user-centred heuristic in place, which, again, depends on the application domain and audience. For example, research suggests [64] that if one of multiple, otherwise equivalent, *time*-ordered events has to be chosen as an explanation, the most recent one will best resonate with the explainee; additionally, prioritising explanations by their *novelty* will keep the explainee engaged and attentive, and distinguishing between *sufficient* and *necessary* conditions for a given outcome can help to reduce cognitive load. While desirable, *brevity* of an explanation can sometimes be at odds with its comprehensiveness and *completeness* – sacrificing the big picture (which in itself may be too convoluted to understand) for concise communication [47]. Explanatory minimalism, nonetheless, bears the danger of oversimplification; however, when it is a strict requirement, explanation *soundness* can be favoured to focus on factors pertinent to the explained instance and discard more general reasons that are largely irrelevant. Such an approach can introduce inaccuracies with respect to the overall black-box system, but the explanations remain truthful for the individual instance. Striking the right balance between generality and specificity of an explanation – as well as achieving all the other aforementioned desiderata – is notoriously challenging and often requires tuning its soundness and completeness for the intended audience and application, which itself may be impractical when done manually and prohibitively difficult through capturing the explainee's mental model.

While posing problems for AI explainers, satisfying this wide range of diverse assumptions and expectations comes naturally for humans when they engage in an **explanatory process** among themselves. This is partly due to shared background knowledge, and is further amplified by interactive communication that allows to rapidly iterate over questions and refine answers to arrive at understanding. One explanation does not fit all [91] and treating explainability as a bi-directional process provides a platform to appreciate uniqueness of each explainee through personalised explanations. While these topics have received relatively little attention in the XAI and IML literature, we can draw design insights and inspirations from research on *explanatory debugging* of predictive models [47]. Therefore, an interactive explanatory process should be *iterative*, enabling the explainee to learn, provide feedback and receive updated insights until reaching a satisfactory conclusion; the explainer ought to always *honour user-provided feedback* by incorporating it into the explanation generation process, or clearly communicate a precise reason if that is impossible; the communication should be *reversible* to allow the explainee to retract a requested change or annul a piece of feedback when it was provided by mistake, or to explore an interesting part of the black box through a speculative enquiry; and, finally, the whole process should be *incremental* to easily attribute each piece of feedback to an explanation change, thereby showing up-to-date results even after small tweaks.

Even though dialogue is fundamental to human explainability, it is largely absent in XAI techniques [91], which are often based on one-way communication, where the user receives a description of the black box without an opportunity to request more details or contest it. A similar interaction in a form of the aforementioned questioning dialogues can also be used to judge the explainee's understanding of the explained concept, thus be a proxy for assessing effectiveness of the explainer. Notably, human dialogue tends to be verbal or written, both of which are based on the natural language. While ubiquitous, this form of communication is not equally effective in conveying all types of information, requiring humans to augment it with visual aids, which are especially helpful when the interaction serves explanatory purposes. The same strategy can be adopted in XAI, where the explainer would switch between various explanatory artefacts – such as (written and spoken) text, images, plots, mathematical formulation, numbers and tables – depending on which one is best suited for the type of information being communicated, i.e., the context. Mixing and matching them is also possible, e.g., a numerical table or a plot complemented with a caption, and may be beneficial as the whole can be greater

than the sum of its parts, especially that certain explanation types may require a specific communication medium or explanatory artefact to be effective. Using visualisation, textualisation and (statistical) summarisation, however, does not guarantee a coherent relation, structure or story conveyed by these communication media alone, which could possibly be achieved by grounding them in a shared context through *logical reasoning* or *formal argumentation* [19].

Contrastive explanations (counterfactuals) dominate the human explanatory process and are considered the gold standard in XAI [64]. They juxtapose a hypothetical situation (foil) next to the factual account with the aim to emphasise the consequences or "would be" change in the outcome. Contrastive statements can be characterised by their lineage: *model-driven* explanations are represented by artificial data points (akin to centroids), whereas *data-driven* explanations are instances recovered from a (training) data set (similar to medoids). Furthermore, the contrast can either be implicit – i.e., "Why class X (and not any other class)?" – or explicit – i.e., "Why class X and not Y?" Counterfactuals are appropriate for lay audiences and domain experts alike, can use concepts of varying difficulty and be expressed in different media such as text and images. They are parsimonious as the foil tends to be based on a single factor, but, if desired, can account for an arbitrary degree of feature covariance. They support interaction, customisation and personalisation, e.g., a foil built around a user-selected feature provided in an explanatory dialogue, which can be used to restrict their search space, possibly making them easier to retrieve. When deployed in a user-centred application, they can provide the explainees with appealing insights by building the foil only around actionable features. However, their effectiveness may be problematic when explaining a black box that is proprietary (e.g., with the intention to protect a trade secret) since contrastive explanations can leak sensitive information, thereby allowing the explainee to steal or game the underlying model. In an open world, they also suffer from vaguely defined or imprecise notions known as *non-concepts* [67], e.g., "What is not-a-dog?"

These idealised properties make contrastive statements appealing, but some may get lost in practice, e.g., an imperfect implementation, resulting in subpar explanations. On the face of it, contrastive explanations resemble causal insights, but unless they are generated with a full causal model [70], they should not be treated as such and instead be interpreted as descriptors of the black box's decision boundary. If they are model-driven, as opposed to data-driven, they may not necessarily come from the data manifold, yielding (out-of-distribution) explanations that are neither feasible nor actionable in the real life, e.g., "Had you been 200 years old, …" Even if they are consistent with the data distribution, the foil may still come from a sparse region, thus prescribing possible but improbable feature values [74]. Contrastive explanations are often specific to a single data point, although humans are known to generalise such insights to unseen and possibly unrelated cases (The Illusion of Explanatory Depth [78]), which may result in overconfidence.

4 THE FALLACY OF SACRIFICING EXPLAINABILITY FOR PREDICTIVE POWER

Theoretical desiderata do not always align with the operationalisation and practicalities of XAI and IML algorithms and the latter are what ends up affecting our lives. For example, explainability is an inherently social process that usually involves bi-directional communication, but most implementations – even the ones using contrastive statements [98, 99] – output a single explanation that is optimised according to some predefined metric, not necessarily addressing concerns of an individual explainee [91]. Similarly, while inherently transparent predictive models and ante-hoc explainers may be preferred [79], such solutions are often model-dependent, labour-intensive and tend to be application-specific, which limits their scope as well as wider applicability and adoption. Instead, post-hoc and model-agnostic explainers dominate the field [58, 76, 77] since they are considered one-stop solutions – a unified explainability experience without a cross-domain adaptation overhead. This silver bullet framework, however, comes at a cost: subpar fidelity that can result in misleading or outright incorrect explanations. While increasingly such considerations find their way into

publications, they are often limited to acknowledging the method's shortcomings, stopping short of offering a viable solution.

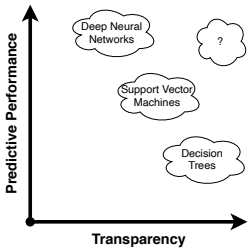

Fig. 1. Fictitious depiction of an anecdotal trade-off between transparency and predictive power of AI systems.

A common belief motivating many methods published in the XAI literature is the perceived *dichotomy* between transparency and predictive power of AI systems. A popular example supporting this theory is the unprecedented effectiveness of deep neural networks on certain ML tasks, whose ever increasing complexity, e.g., the number of layers and hidden units, improves their performance at the expense of transparency. This trade-off has been reiterated in the DARPA XAI program's Broad Agency Announcement [32] and supported by an appealing graph reproduced in Figure 1. However, it is a *theory* based mostly on anecdotal evidence, with Rudin [79] criticising plots like this given their lack of scale, transparency or precise performance metrics, and supporting data. Notably, Rudin argues that investing more effort into feature engineering can help to build inherently explainable AI systems that perform on a par with their black-box alternatives [16].

This anecdotal trade-off and a tendency to prioritise predictive power mean that explainability is often only an afterthought. Such a mindset contributes to an AI landscape with an abundance of well-performing but inherently opaque algorithms that are in need of explainability, thus creating a demand for universal explainers that are post-hoc and model-agnostic, such as surrogates [17, 76]. This seemingly uncompromising development approach – where state-of-the-art performance remains the main objective, later complemented with a post-hoc explainer – offers an attractive alternative (and rebuttal) to designing inherently explainable AI system, whose creation arguably requires more effort. While such explainers are compatible with any black-box model, they are not necessarily equally well suited for all of them – after all the computer science folklore of "no free lunch" (a single, universal algorithm cannot outperform all the others across the board) applies here as well. Some post-hoc and model-agnostic explainers boast appealing properties and guarantees, however upon closer inspection one often encounters caveats and assumptions required for these to hold, such as the underlying "black box" being a linear model [58]. Making an explainer model-agnostic introduces an extra layer of complexity that usually entails a degree of randomness and decreased fidelity [96, 106], so that using them becomes a stopgap to claim explainability of an inherently opaque AI system rather than addressing genuine explainability needs.

In Rudin's [79] view, many high-stakes AI systems can be made explainable by design with enough effort put towards data pre-processing and feature engineering (which otherwise, e.g., for neural networks, may go into architecture search and parameter tuning). Such ante-hoc explainers are usually domain-specific and after the initial engineering endeavour they are easy to manage and maintain. While this approach should be championed for structured (tabular) data where it has been shown to perform on a par with state-of-the-art black boxes [16], the same may be unachievable for sensory data such as images and sounds, for which opaque models, e.g., deep neural networks, have the upper hand. In addition to black boxes modelling sensory data, pre-existing, inaccessible or legacy AI systems may require interpretability, in which case they can only be retrofitted with post-hoc explainers. However, falling back on off-the-shelf solutions may not guarantee acceptable fidelity [96] (in particular, soundness and completeness), which is of particular importance and may require tailor-made explainers and transparent communication of their limitations.

While developing a predictive pipeline, we have an abundance of pre-processing and modelling tools and techniques at our disposal, a selection of which will end up in the final system. The XAI and IML landscape, on the other hand, is quite different, especially for post-hoc and model-agnostic approaches: explainers tend to be end-to-end tools with

only a handful of parameters exposed to the user. In view of the no free lunch theorem, this is undesirable as despite being model-agnostic, i.e., compatible with any model type, these monolithic algorithms cannot perform equally well for every one of them [96]. This variability in their behaviour can often be attributed to a misalignment between the assumptions baked into an explainer and the properties of the explained system, which manifests itself in low fidelity.

Model-specific or ante-hoc explainers as advocated by Rudin [79] can be used to address this issue; however, as discussed earlier, such a solution may have limited applicability and cannot be retrofitted to pre-existing AI systems. Resolving a similar challenge in machine learning and data mining often comes down to a series of investigative steps to guide algorithmic choices down the line, which can be operationalised within a standardised process for knowledge discovery such as KDD [20], CRISP-DM [15, 61] or BigData [2]. For example, by analysing feature correlation, data uniformity and class imbalance, we can account for these phenomena when engineering features and training models, thereby making the resulting AI systems more accountable and robust. Nonetheless, while we may have a set of universal properties expected of XAI and IML systems [89], we lack a dedicated process that could guide the development and assessment of explainers – their practical requirements and needs – which likely hinders adherence to best practice. Although one can imagine a generic workflow for designing inherently interpretable (ante-hoc) systems [79], a similar endeavour should not be neglected for model-agnostic and post-hoc explainers that could be adapted to individual predictive black boxes by capitalising on their flexibility and modularity [96], possibly overcoming low fidelity [90].

While some researchers claim that we should not expect machine learning algorithms, such as deep neural networks, to be explainable and instead regulate them purely based on their real-life performance [85], it is not a widely shared belief [37]. This insight comes from the alleged inability of humans to explain their actions since such justifications are post-factum stories that are concocted and retrofitted for the benefit of the audience. Certifying autonomous agents based on their output, on the other hand, is consistent with human values as one can hypothesise about committing a crime, but one cannot be punished unless such a thought is acted upon. While the origin and nature of human thought processes may be shrouded in mystery, its formulation is expected to follow the reason of logic to be (socially) acceptable. In particular, Miller [63] refutes performance-based validation by arguing that explainability stemming from regulatory requirements is secondary to concerns arising from societal values such as ethics and trust. Importantly, an appropriate and comprehensive explainability solution can also become a technological springboard to reducing or eliminating bias [81, 82], unfairness [14, 49, 68] and lack of accountability (to the benefit of robustness [1, 29], safety [4, 30] and security) from data-driven predictive models, thus improving their trustworthiness [87].

5 EXPLANATION DIVERSITY AND MULTIPLICITY: WHAT TO EXPLAIN AND HOW TO EXPLAIN IT

So far we have primarily focused on explaining predictions and actions of intelligent systems since they are observable and can be related to by a wide range of explainees regardless of their background. However, automated **predictions** are just artefacts of a more elaborate artificial intelligence or machine learning predictive process, which manipulates **data** to learn **models** that generalise well, thus are capable of predicting (previously unseen) instances [23]. Since any element of this workflow can be opaque [93, 94], comprehensive explanations may need to consist of insights pertaining to the entire predictive pipeline, discussing diverse topics such as data collection and (pre-)processing, modelling caveats and assumptions, and the meaning and interpretation of predictions, all of which can be bundled together in a shared user interface to provide a multi-faceted view of the investigated system [45, 46, 104]. Additionally, as each explanation may just provide a small, and quite possibly distorted, reflection of the true behaviour of a black box, achieving the desired level of transparency (and understanding) can require communicating multiple, complementary insights for each unintelligible step or observation, which in turn bears the danger of overwhelming and confusing the explainee.

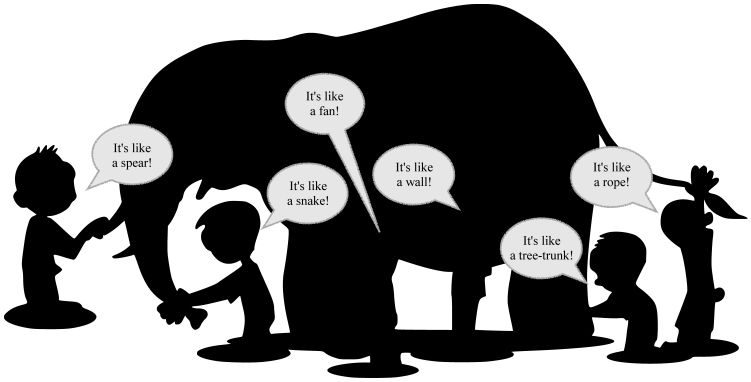

Fig. 2. Depiction of "The Blind Men and the Elephant" parable [80] illustrating that any complex subject can be studied in many ways. It also symbolises that individual pieces of evidence may often be contradictory and insufficient to understand the bigger picture without first being aggregated and grounded within a shared context.

This multitude of explanatory information has to be navigated carefully and can be understood as unique probing and inspection techniques that without a shared context may yield competing or even contradictory evidence akin to the parable of "The Blind Men and the Elephant" [80] – see Figure 2. Furthermore, as AI and ML processes are directional – from data, through models, to predictions – the latter components depend on the former, which also applies to their respective explanations. For example, if data attributes are incomprehensible, explanations of models and predictions expressed in terms of these features will also be opaque.

Explaining data may be challenging without any modelling assumptions, hence there may not necessarily exist a pure data explanation method beyond simple *summary statistics* (e.g., class ratio, per-class feature distribution) and *descriptors* (e.g., "the classes are balanced", "the data are bimodal", "these features are highly correlated"). Note that the former simply state well-defined properties and may not be considered explanations, whereas the latter can be contrastive and lead to understanding. Importantly, data are already a model – they express a (subjective and partial) view of a phenomenon and come with certain assumptions, measurement errors or even embedded cultural biases (e.g., "How much is a lot?"). "Data statements" [8], "data sheets" [26] and "nutrition labels" [35] attempt to address such concerns by capturing these (often implicit) assumptions. As a form of data explanations, they characterise important aspects of data and their collection process in a coherent way, e.g., experimental setup, collection methodology (by whom and for what purpose), pre-processing (cleaning and aggregation), privacy aspects, the data owners, and so on.

Explaining models in whole or in parts (e.g., specific sub-spaces or cohorts) should engender a general, truthful and accurate understanding of their functionality. While some models may be inherently transparent, e.g., shallow decision trees, their simulatability [55] – the explainee's ability to simulate their decision process mentally *in vivo* – may not produce understanding (see Section 2). Popular model explanations include feature importance [12, 21], feature influence on predictions [24], presenting the model in cognitively-digestible portions [45, 86] and model simplification [17] (e.g., mimicking its behaviour or a global surrogate). Since not all models operate directly on the input features, an *interpretable representation* may be necessary to convey an explanation, e.g., a super-pixel segmentation of an image [76]; alternatively, if the data are comprehensible, landmark exemplars can be used to explain the behaviour of a model or its parts [39, 40].

Predictions are explained to communicate a rationale behind a particular decision of a model. Depending on the explanation type, a range of diverse aspects concerning the model's decisive process can be provided to the explainee. For example, the user may be interested in feature importance [76], feature influence [58], relevant data examples [38] and training instances [44], or contrastive statements [74, 99], to name a few. Note that while some of these explanation types are similar to model explanations, here they are explicitly generated with respect to a single data point and may not necessarily generalise beyond this particular case, whereas for model explanations they convey similar information for all data (i.e., the model). A good example of this duality is information communicated by Individual Conditional Expectation [28] (ICE) and Partial Dependence [24] (PD), both of which are feature influence explanations – the first with respect to a single data point and the latter concerning a model – as shown in Figure 3. Akin to model explanations, the information can be conveyed in the raw feature space or using an interpretable representation.

With such a diverse range (and possibly large quantity) of explanations, their presentation requirements – **communication media** and **protocols** [89] – will naturally vary [93, 94]. A simple approach to characterise an AI component is (statistical) *summarisation* – it is commonly used for describing properties of data with numerical tables and vectors, which can be difficult to digest for non-experts. *Visualisation* – a graphical representation of a phenomenon – is a more advanced, insightful and flexible analytical tool. Static figures communicate information in one direction, similar to summarisation; however, creating interactive plots can facilitate a "dialogue" with an explainee, thereby catering to a more diverse audience. Visualisations are often supported by a short narrative in the form of a caption, which increases their informativeness. *Textualisation* – a natural language description of a phenomenon – can express concepts of higher complexity and dimensionality than plots, which can help to overcome the curse of dimensionality and the inherent limitations of the human visual system. Communicating with text enables a true dialogue and has been shown to be more insightful and effective than presenting raw, numerical and visual data [73], which can accompany the narrative to improve its expressiveness. A further refinement of textualisation is formal *argumentation* [19] – a structured and logically-coherent dialogue accounting for every disputable statement and giving the explainee an opportunity to contest the narrative, thus providing explanations leading to understanding rather than informative descriptions.

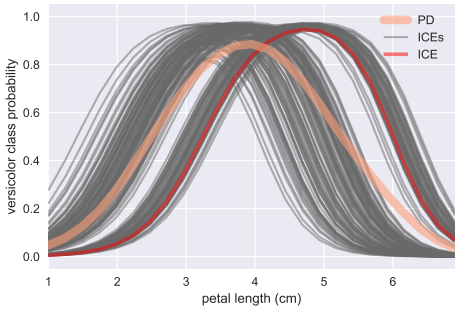

Fig. 3. Explanation of a model predicting the probability of the *versicolor* class when varying the *petal length* attribute for the Iris data set [22]. Individual Conditional Expectation of a selected instance is plotted in red, and the orange curve is the Partial Dependence of the model computed by averaging all individual ICEs (in grey).

Thus far we have been mainly concerned with AI and ML explainability on a relatively abstract level, all of which constitute just a small portion of XAI and IML research. In an ideal world, relevant publications would consider many of the aforementioned factors and build their mechanics around them, however it has only recently become a trend and numerous early pieces of work lack such a reflection. To complement the viewpoint presented in the preceding sections and bridge the *theoretical* (foundational and social) and *technical* (algorithmic and engineering) aspects of explainers we traverse through **practical explainability research**. Without aiming to be exhaustive – given the availability of several comprehensive surveys [31, 54] – we finish this section by identifying a number of landmark contributions that have influenced the entire research field. We also omit topics adjacent to explainability, such as interactive exploratory user interfaces [36, 105], creative visualisations of explainability approaches [46] and systems combining multiple explainability techniques within a single tool [104].

The most popular explainers are *model-agnostic* and *post-hoc* as they can be retrofitted into any predictive black box (at the expense of adding a modelling layer that may negatively impact explanation fidelity). These include RuleFit [25], Local Interpretable Model-agnostic Explanations (LIME [76]), anchors [77], SHapley Additive exPlanations (SHAP [58]), Black-box Explanations through Transparent Approximations (BETA [50, 51]), PD [24], ICE [28] and Permutation Importance (PI [12]), among many others. Most of these methods operate directly on raw data, with the exception of LIME and anchors, which use interpretable representations to improve intelligibility of explanations composed for complex data domains such as text and images. Another attractive avenue of explainability research, which partly overlaps with post-hoc methods, is opening up (deep) neural networks by designing tools and techniques *specific to these models* or, more broadly, compatible with differentiable predictors. These models are notoriously opaque, however their superior predictive performance for a wide spectrum of applications accelerates their popularity and widespread adoption [52]. Relevant explainability techniques include global surrogates [17], saliency maps [107], influential training instances [44], counterfactuals [99] (which are surprisingly similar to the problematic adversarial examples [29]), and influential high-level, human-intelligible insights based on Testing with Concept Activation Vectors (TCAV [41]). An alternative XAI and IML research agenda concentrates on *inherently explainable* predictive models, and *ante-hoc* explainers designed for popular black boxes. Examples of the former are generalised additive models [57] and falling rule list [103]; whereas the latter include global and local explanations of naïve Bayes classifiers [47], and clustering insights based on prominent exemplars and dominating features [40].

6 TOWARDS INTELLIGIBLE AND ROBUST EXPLAINERS

In this paper we explored the relatively recent and still evolving fields of artificial intelligence explainability and machine learning interpretability. We introduced the main topics and provided the philosophical, theoretical and technical background needed to appreciate the depth and complexity of this research. In particular, we highlighted two different mental models: *functional* – enough understanding to operationalise a concept; and *structural* – in-depth, theoretical appreciation of underlying processes. We further argued that the former – a shallow form of understanding – aligns with The Chinese Room Argument [72, 83] and the notion of simulatability [55]. We also reviewed diverse notions of explainability, interpretability, transparency, intelligibility and many others that are often used interchangeably in the literature, and argued in favour of *explainability*. We defined this concept as (logical) *reasoning* applied to transparent XAI and IML insights interpreted under specific *background knowledge* – a process that engenders *understanding* in explainees. We used these observations to challenge the popular view that decision trees are explainable just because they are transparent. Deep or wide trees lack interpretability, which can be restored by applying a suitable form of logical reasoning – a prerequisite of explainability – undertaken by either an algorithm or a human investigator.

While the most visible aspect of XAI and IML research is the technology that enables it, explainees – the recipients of such explanations who tend to be humans – are just as important (and ought to be treated as first-class citizens) since their *understanding* of the underlying predictive system determines the ultimate success of an explainer. We explored this topic by looking at human-centred explainability and various desiderata that this concept entails, in particular focusing on explicitly acknowledging presence of humans and projecting the explanations directly at them. To this end, we pursued important insights from the social sciences that prescribe how to adapt machine explainability to fulfil expectations of the explainees, hence achieve seamless explanatory interaction. The two crucial observations in this space are: (i) a preference for (meaningful) *contrastive* explanations, which form the cornerstone of human-centred explainability; and (ii) facilitating an interactive, dialogue-like, bi-directional explanatory *process* – akin to a conversation – as opposed to delivering a one-off "take it or leave it" explanation to ensure coherence with people's expectations

regardless of their background knowledge and prior experience with this sort of technology. Notably, the explanation type and delivery medium should also be adapted to the circumstances. This is particularly important when the audience is diverse as one predefined type of an explanation may be insufficient since it is unlikely to address all the possible questions and unique perspectives. An XAI explainer that communicates through contrastive explanations and provides the explainees with an opportunity to interactively customise and personalise them [59] – offering a chance to contest and rebut them in case of a disagreement – should therefore be considered the gold standard [64, 99].

In addition to enhancing explainee satisfaction, operating within this purview has other, far-reaching benefits such as enabling algorithmic fairness evaluation, accountability assessment and debugging of predictive models. It is also compatible with all the elements of the machine learning workflow – which consists of data, models and predictions – as each of these components may be in need of interpretability. In view of a variety of explainability approaches, each operating in a unique way, we also looked at the disputed trade-off between explainability and predictive power, the existence of which has only been supported by anecdotal evidence thus far. We then connected this debate to the distinction between inherent (ante-hoc) and retrofitted (post-hoc) explainability: the former provides explanations of superior quality but requires extensive engineering effort to be built, whereas the latter is flexible and universal at the expense of fidelity. While the former may be shunned due to the required effort, we argued that building trustworthy post-hoc explainers may be just as complicated and require just as much commitment since these seemingly easy to use tools conceal a complex process governing their composition and influencing their quality behind the facade of universality [90, 92, 96, 97]. This considerable effort required to set them up, therefore, illuminates a crucial question: Is it better to spend time and effort on configuring post-hoc explainers or instead invest these resources into building inherently explainable predictive models? Unsurprisingly, there is no definitive answer given the uniqueness of each individual case, e.g., legacy systems and predictors built from scratch.

Regardless of the particular implementation and operationalisation details, explainers of black boxes should adopt and embody as many of these findings as possible to engender trust in data-driven predictive systems. Since each explanation reveals just a fragment of the black box and only the right mixture of evidence can paint the full picture, XAI and IML approaches need to be responsive and adapt seamlessly to the user's requests and expectations. Such an engaging algorithmic interlocutor should build logically consistent narratives and serve more as a guide and a teacher than a facts reporter. To this end, we need to develop an explanatory process built on top of a system that enables logical reasoning between intelligent agents: human–machine or machine–machine. An appropriate foundation – managing the dialogue as well as tracking and storing the evolving knowledge base of the involved parties – should benefit and encourage an interdisciplinary research agenda drawing from multiple areas of computer and social sciences. In the end, nonetheless, the explainee needs to be a savvy interrogator, asking the right questions and firmly navigating the entire process to understand the behaviour of such data-driven oracles. After all, in Arthur C. Clarke's words: "Any sufficiently advanced technology is indistinguishable from magic." While this view may partially reflect a broader perception of artificial intelligence and machine learning applications, this paper reconciles XAI and IML research published to date to establish a solid foundation for addressing open questions in an effort to demystify predictive algorithms.

ACKNOWLEDGMENTS

This research was supported by the ARC Centre of Excellence for Automated Decision-Making and Society, funded by the Australian Government through the Australian Research Council (project number CE200100005); and the TAILOR project, funded by EU Horizon 2020 research and innovation programme under GA No 952215.

REFERENCES

[1] Evan Ackerman. 2019. Three small stickers in intersection can cause Tesla autopilot to swerve into wrong lane. *IEEE Spectrum, April* 1 (2019).

[2] Divyakant Agrawal, Philip Bernstein, Bertino Elisa, Davidson Susan, Dayal Umeshwar, Michael Franklin, and Y Papakonstantinou. 2012. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium. *Committee of the Computing Research Association* (2012).

[3] David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Weight of evidence as a basis for human-oriented explanations. *2019 Workshop on Human-Centric Machine Learning (HCML 2019) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada* (2019). https://arxiv.org/abs/1910.13503 arXiv preprint arXiv:1910.13503.

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica, May* 23 (2016), 2016.

[5] Abdallah Arioua and Madalina Croitoru. 2015. Formalizing explanatory dialogues. In *International Conference on Scalable Uncertainty Management*. Springer, 282–297.

[6] Roy F Baumeister and Leonard S Newman. 1994. Self-regulation of cognitive inference and decision processes. *Personality and Social Psychology Bulletin* 20, 1 (1994), 3–19.

[7] Boris Beizer. 1995. *Black-box testing: Techniques for functional testing of software and systems*. John Wiley & Sons, Inc.

[8] Emily M Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[9] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[10] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2017), Melbourne, Australia*.

[11] Or Biran and Kathleen McKeown. 2014. Justification narratives for individual classifications. In *Proceedings of the AutoML workshop at ICML*, Vol. 2014. 1–7.

[12] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[13] Mario Bunge. 1963. A general black box theory. *Philosophy of Science* 30, 4 (1963), 346–358.

[14] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, New York, NY, USA, 77–91. http://proceedings.mlr.press/v81/buolamwini18a.html

[15] Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rudiger Wirth, et al. 2000. CRISP-DM 1.0: Step-by-step data mining guide. *SPSS inc* 9 (2000), 13.

[16] Chaofan Chen, Kangcheng Lin, Cynthia Rudin, Yaron Shaposhnik, Sijia Wang, and Tong Wang. 2018. An interpretable model with globally consistent explanations for credit risk. *2018 Workshop on Challenges and Opportunities for AI in Financial Services: the Impact of Fairness, Explainability, Accuracy, and Privacy at the 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada* (2018). https://arxiv.org/abs/1811.12615 arXiv preprint arXiv:1811.12615.

[17] Mark Craven and Jude W Shavlik. 1996. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*. 24–30.

[18] Richard Dawkins. 2011. The tyranny of the discontinuous mind. *New Statesman* 19 (2011), 54–57.

[19] Phan Minh Dung, Robert A Kowalski, and Francesca Toni. 2009. Assumption-based argumentation. In *Argumentation in artificial intelligence*. Springer, 199–218.

[20] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. From data mining to knowledge discovery in databases. *AI magazine* 17, 3 (1996), 37–37.

[21] Aaron Fisher, 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, 177 (2019), 1–81.

[22] Ronald A Fisher. 1936. The use of multiple measurements in taxonomic problems. *Annals of eugenics* 7, 2 (1936), 179–188.

[23] Peter Flach. 2012. *Machine Learning: The art and science of algorithms that make sense of data*. Cambridge University Press.

[24] Jerome H Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[25] Jerome H Friedman, Bogdan E Popescu, et al. 2008. Predictive learning via rule ensembles. *The Annals of Applied Statistics* 2, 3 (2008), 916–954.

[26] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2018) at the 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden* (2018). https://arxiv.org/abs/1803.09010 arXiv preprint arXiv:1803.09010.

[27] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining Explanations: An Overview of Interpretability of Machine Learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 80–89.

[28] Alex Goldstein, 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, 1 (2015), 44–65.

[29] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. *Proceedings of the 3rd International Conference on Learning Representations (ICLR), San Diego, California* (2015). http://arxiv.org/abs/1412.6572 arXiv preprint arXiv:1412.6572.

[30] Adam Grzywaczewski. 2017. Training AI for self-driving vehicles: The challenge of scale. *NVIDIA Developer Blog, October* (2017).

[31] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[32] David Gunning. 2016. *Broad Agency Announcement, Explainable Artificial Intelligence (XAI).* Technical Report. Defense Advanced Research Projects Agency (DARPA). https://www.darpa.mil/attachments/DARPA-BAA-16-53.pdf

[33] David Gunning. 2017. Explainable Artificial Intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)* 2 (2017), 2.

[34] Mark Hall, Daniel Harborne, Richard Tomsett, Vedran Galetic, Santiago Quintana-Amate, Alistair Nottle, and Alun Preece. 2019. A Systematic Method to Understand Requirements for Explainable AI (XAI) Systems. In *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2019), Macau, China.*

[35] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv preprint arXiv:1805.03677* (2018). https://arxiv.org/abs/1805.03677

[36] Google Inc. 2015. TensorBoard: TensorFlow's visualization toolkit. https://www.tensorflow.org/tensorboard

[37] Hessie Jones. 2018. Geoff Hinton Dismissed The Need For Explainable AI: 8 Experts Explain Why He's Wrong.

[38] Been Kim, Elena Glassman, Brittney Johnson, and Julie Shah. 2015. iBCM: Interactive Bayesian case model empowering humans via intuitive interaction. *MIT Libraries Technical Report: MIT-CSAIL-TR-2015-010* (2015).

[39] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for interpretability. In *Advances in Neural Information Processing Systems*. 2280–2288.

[40] Been Kim, Cynthia Rudin, and Julie A Shah. 2014. The Bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*. 1952–1960.

[41] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*. 2668–2677.

[42] Alexandra Kirsch. 2017. Explain to whom? Putting the User in the Center of Explainable AI. In *Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 colocated with 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy.*

[43] Derek J Koehler. 1991. Explanation, imagination, and confidence in judgment. *Psychological bulletin* 110, 3 (1991), 499.

[44] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 1885–1894. http://proceedings.mlr.press/v70/koh17a.html

[45] Josua Krause, Adam Perer, and Enrico Bertini. 2016. Using visual analytics to interpret predictive machine learning models. *Workshop on Human Interpretability in Machine Learning (WHI 2016) at the 33rd International Conference on Machine Learning (ICML 2016), New York, New York* (2016). https://arxiv.org/abs/1606.05685 arXiv preprint arXiv:1606.05685.

[46] Josua Krause, 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*. ACM, 5686–5697.

[47] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, 126–137.

[48] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Computing (VL/HCC), 2013 IEEE Symposium on*. IEEE, 3–10.

[49] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[50] Himabindu Lakkaraju, 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*. 1675–1684.

[51] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017) at the 23rd SIGKDD conference on Knowledge Discovery and Data Mining (KDD 2017), Halifax, Nova Scotia, Canada* (2017). https://arxiv.org/abs/1707.01154 arXiv preprint arXiv:1707.01154.

[52] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436.

[53] Cornelius T Leondes. 2001. *Expert systems: The technology of knowledge management and decision making for the 21st century.* Elsevier.

[54] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. 2021. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* 23, 1 (2021), 18.

[55] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 16, 3, Article 30 (jun 2018), 27 pages. https://doi.org/10.1145/3236386.3241340

[56] Tania Lombrozo. 2006. The structure and function of explanations. *Trends in cognitive sciences* 10, 10 (2006), 464–470.

[57] Yin Lou, 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*. ACM, 623–631.

[58] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[59] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1033–1041.

[60] David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press.

[61] Fernando Martínez-Plumed, Lidia Contreras-Ochando, Cèsar Ferri, José Hernández Orallo, Meelis Kull, Nicolas Lachiche, Maréa José Ramírez Quintana, and Peter A Flach. 2019. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering* (2019).

[62] R Mead. 2015. When a teacher's job depends on a child's test. *The New Yorker* (2015).

[63] Tim Miller. 2019. "But why?" Understanding explainable artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 20–25.

[64] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[65] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. In *Proceedings of the IJCAI Workshop on eXplainable Artificial Intelligence (XAI 2017), Melbourne, Australia*. https://arxiv.org/abs/1712.00547 arXiv preprint arXiv:1712.00547.

[66] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[67] Fabian Offert. 2017. "I know it when I see it". Visualization and Intuitive Interpretability. *2017 Symposium on Interpretable Machine Learning at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California* (2017). http://arxiv.org/abs/1711.08042 arXiv preprint arXiv:1711.08042.

[68] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2019. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data* 2 (2019), 13.

[69] Cathy O'neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

[70] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

[71] Judea Pearl and Dana Mackenzie. 2018. *The book of why: The new science of cause and effect*. Basic Books.

[72] Roger Penrose. 1989. *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

[73] François Portet, Ehud Reiter, Albert Gatt, Jim Hunter, Somayajulu Sripada, Yvonne Freer, and Cindy Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence* 173, 7 (2009), 789–816.

[74] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. 2020. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 344–350.

[75] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *Proceedings of the AAAI Fall Symposium on Artificial Intelligence in Government and Public Sector, Arlington, Virginia, USA* (2018). arXiv preprint arXiv:1810.00184.

[76] Marco Tulio Ribeiro, 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, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[77] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[78] Leonid Rozenblit and Frank Keil. 2002. The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive science* 26, 5 (2002), 521–562.

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

[80] John G Saxe. 2016. *The blind men and the elephant*. Enrich Spot Limited.

[81] Nripsuta Ani Saxena. 2019. Perceptions of Fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 537–538.

[82] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[83] John R Searle. 1980. Minds, brains, and programs. *Behavioral and brain sciences* 3, 3 (1980), 417–424.

[84] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.

[85] Tom Simonite. 2019. Google's AI guru wants computers to think more like brains.

[86] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B Viégas, and Martin Wattenberg. 2016. Embedding projector: Interactive visualization and interpretation of embeddings. *Workshop on Interpretable Machine Learning in Complex Systems at the 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain* (2016). https://arxiv.org/abs/1611.05469 arXiv preprint arXiv:1611.05469.

[87] Kacper Sokol. 2019. Fairness, Accountability and Transparency in Artificial Intelligence: A Case Study of Logical Predictive Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 541–542.

[88] Kacper Sokol and Peter Flach. 2019. Desiderata for Interpretability: Explaining Decision Tree Predictions with Counterfactuals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 10035–10036.

[89] Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 56–67.

[90] Kacper Sokol and Peter Flach. 2020. LIMEtree: Interactively Customisable Explanations Based on Local Surrogate Multi-output Regression Trees. (2020). https://arxiv.org/abs/2005.01427 arXiv preprint arXiv:2005.01427.

[91] Kacper Sokol and Peter Flach. 2020. One explanation does not fit all. *KI-Künstliche Intelligenz* (2020), 1–16.

[92] Kacper Sokol and Peter Flach. 2020. Towards Faithful and Meaningful Interpretable Representations. *arXiv preprint arXiv:2008.07007* (2020). http://arxiv.org/abs/2008.07007

[93] Kacper Sokol and Peter A Flach. 2017. The Role of Textualisation and Argumentation in Understanding the Machine Learning Process. In *IJCAI*. 5211–5212.

[94] Kacper Sokol and Peter A Flach. 2017. The Role of Textualisation and Argumentation in Understanding the Machine Learning Process: A position paper. In *Automated Reasoning Workshop*. 11–12.

[95] Kacper Sokol and Peter A Flach. 2018. Glass-Box: Explaining AI Decisions With Counterfactual Statements Through Conversation With a Voice-enabled Virtual Assistant. In *IJCAI*. 5868–5870.

[96] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. 2019. bLIMEy: Surrogate Prediction Explanations Beyond LIME. *2019 Workshop on Human-Centric Machine Learning (HCML 2019) at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada* (2019). https://arxiv.org/abs/1910.13016 arXiv preprint arXiv:1910.13016.

[97] Kacper Sokol, Alexander Hepburn, Raul Santos-Rodriguez, and Peter Flach. 2020. What and How of Machine Learning Transparency: Building Bespoke Explainability Tools with Interoperable Algorithmic Components. *Hands-on Tutorial at The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML-PKDD), Ghent, Belgium* (2020). https://events.fat-forensics.org/2020_ecml-pkdd

[98] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. 2018. Contrastive explanations with local foil trees. *Workshop on Human Interpretability in Machine Learning (WHI 2018) at the 35th International Conference on Machine Learning (ICML 2018), Stockholm, Sweden* (2018). https://arxiv.org/abs/1806.07470 arXiv preprint arXiv:1806.07470.

[99] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GPDR. *Harv. JL & Tech.* 31 (2017), 841.

[100] Douglas Walton. 2007. Dialogical Models of Explanation. *ExaCt* 2007 (2007), 1–9.

[101] Douglas Walton. 2011. A dialogue system specification for explanation. *Synthese* 182, 3 (2011), 349–374.

[102] Douglas Walton. 2016. A dialogue system for evaluating explanations. In *Argument Evaluation and Evidence*. Springer, 69–116.

[103] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial Intelligence and Statistics*. 1013–1022.

[104] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[105] James Wexler. 2017. Facets: An open source visualization tool for machine learning training data. *Google Open Source Blog* (2017).

[106] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. "Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations. *AI for Social Good Workshop at the 36th International Conference on Machine Learning (ICML 2019), Long Beach, California* (2019). https://arxiv.org/abs/1904.12991 arXiv preprint arXiv:1904.12991.

[107] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *Proceedings of the 5th International Conference on Learning Representations (ICLR), Toulon, France* (2017). https://openreview.net/forum?id=BJ5UeU9xx arXiv preprint arXiv:1702.04595.