NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
Amy N. Yates
Kristen Greene
David A. Broniatowski
Mark A. Przybocki

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8312

National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
Amy N. Yates
Kristen Greene
Information Access Division
Information Technology Laboratory

David A. Broniatowski
Information Technology Laboratory

Mark A. Przybocki
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8312

September 2021

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

National Institute of Standards and Technology
Interagency or Internal Report 8312
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8312, 43 pages (September 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8312

Abstract

We introduce four principles for explainable artificial intelligence (AI) that comprise fundamental properties for explainable AI systems. We propose that explainable AI systems deliver accompanying evidence or reasons for outcomes and processes; provide explanations that are understandable to individual users; provide explanations that correctly reflect the system's process for generating the output; and that a system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output. We have termed these four principles as explanation, meaningful, explanation accuracy, and knowledge limits, respectively. Through significant stakeholder engagement, these four principles were developed to encompass the multidisciplinary nature of explainable AI, including the fields of computer science, engineering, and psychology. Because one-size-fits-all explanations do not exist, different users will require different types of explanations. We present five categories of explanation and summarize theories of explainable AI. We give an overview of the algorithms in the field that cover the major classes of explainable algorithms. As a baseline comparison, we assess how well explanations provided by people follow our four principles. This assessment provides insights to the challenges of designing explainable AI systems.

Key words

Artificial Intelligence (AI); explainable AI; explainability; trustworthy AI.

Executive Summary

The AI space is vast, complicated, and continually evolving. With advances in computing power and ever-larger datasets, AI algorithms are being explored and developed for use in a wide variety of application spaces, with a variety of potential users and associated risks. The AI community is pursuing explainability as one of many desirable characteristics for trustworthy AI systems. Working with the AI community, NIST has identified additional technical characteristics to cultivate trust in AI. In addition to explainability and interpretability, among other AI system characteristics proposed to support system trustworthiness are accuracy, privacy, reliability, robustness, safety, security (resilience), mitigation of harmful bias, transparency, fairness, and accountability. Explainability and other AI system characteristics interact at various stages in the AI lifecycle. While all are critically important, this work focuses solely on principles of explainable AI systems.

In this paper, we introduce four principles that we believe comprise fundamental properties for explainable AI systems. These principles of explainable AI were informed by engagement with the larger AI community through a NIST public workshop and public comment period. We recognize that not all AI systems may require explanations. However, for those AI systems that are intended or required to be explainable, we propose that those systems adhere to the following four principles:

Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.

Meaningful: A system provides explanations that are understandable to the intended consumer(s).

Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.

Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

In this work, we recognize the importance of both process-based and outcome-based explanations, as well as the importance of explanation purpose and style. For example, AI developers and designers may have very different explanation needs than policy makers and end users. Therefore, why an explanation is requested and how it is delivered may differ depending on the AI users. These four principles are heavily influenced by considering the AI system's interaction with the human recipient of the information. The requirements of the given situation, the task at hand, and the consumer will all influence the type of explanation deemed appropriate for the situation. These situations can include, but are not limited to, regulator and legal requirements, quality control of an AI system, and customer relations. Our four principles of explainable AI systems are intended to capture a broad set of motivations, reasons, and perspectives. The principles allow for defining the contextual factors to consider for an explanation, and pave the way forward to measuring explanation quality.

We imagine that given the complexity of the AI space, these principles will benefit from additional refinement and community input over time. We fully acknowledge that there are numerous other socio-technical factors that influence AI trustworthiness beyond explainability. This work on principles of explainable AI systems is part of the much larger NIST AI portfolio[1] around trustworthy AI data, standards, evaluation, validation, and verification—all necessary for AI measurements. NIST is a metrology institute and as such, defining initial principles of explainable AI systems acts as a roadmap for future measurement and evaluation activities. The agency's AI goals and activities are prioritized and informed by its statutory mandates, White House directions, and the needs expressed by U.S. industry, other federal agencies, and the global AI research community. The current work is but one step in this much larger space, and we imagine this work will continue to evolve and progress over time, much like the larger AI field.

[1]NIST AI Program Fact Sheet: https://www.nist.gov/system/files/documents/2021/08/10/AI%20Fact%20Sheet%200615%20FINAL.pdf

Table of Contents

List of Figures

1. Introduction

When the father of one of the authors was diagnosed with cancer, they went to speak with his oncologist. The oncologist described the state of his cancer and went through strategies and options for treatment. The oncologist answered the father's questions and explained his role in his treatment. The author's father felt he was a partner and had some control. The father trusted the treatment because he received a meaningful and understandable explanation about the process. The doctor's bedside manner won over the father. The medical arts have changed, and possessing a good bedside manner has become de rigueur. When artificial intelligence (AI) systems contribute to a diagnosis, they could support good bedside manners by explaining their recommendations to physicians.

Medical diagnoses are just one example where AI systems contribute to decisions that impact a person's life. Other examples are systems which evaluate loan applications and recommend jail sentences. The nature of these decisions has spurred a drive to create algorithms, methods, and techniques to accompany outputs from AI systems with explanations. This drive is motivated in part by laws and regulations which state that decisions, including those from automated systems, must provide information about the reasoning behind those decisions[2]. It is also motivated by the desire to create trustworthy AI [49, 109, 131].

Explainable AI is one of several properties that characterize trust in AI systems [121, 127, 134]. Other properties include resiliency, reliability, bias, and accountability. Usually, these terms are not defined in isolation, but as a part or set of principles or pillars. The definitions vary by author, and they focus on the norms that society expects AI systems to follow. Based on the calls for explainable systems [59], it can be assumed that the failure to articulate the rationale for an answer can affect the level of trust users will grant that system. Suspicions that the system is biased or unfair can raise concerns about harm to individuals and to society [119, 146]. This may slow societal acceptance and adoption of the technology.

With the increased call for explanations, the field needs a principled method that characterizes a good explanation from an AI system. First, the characterization needs to be human-centered, because humans consume them. Second, they need to be understandable to people. Third, explanations should correctly reflect the system's process for generating the output. To foster confidence in explanations, the system should indicate when it is operating outside its designed conditions. These core concepts of a good explanation are the basis for our four principles of explainable AI.

Although these principles may affect the methods in which algorithms operate to meet explainable AI goals, the focus of the concepts is not algorithmic methods or computations themselves. Also, the principles do not pertain to the system's usage during deployment. Rather, we present four principles organized around the humans that consume the explanations. They provide a structure to begin measuring components of explanations: their quality, goodness, accuracy, and limitations. To measure explanations in a structured way

[2]The Fair Credit Reporting Act (FCRA) and the European Union (E.U.) General Data Protection Regulation (GDPR) Article 13.

is essential for the field to make progress toward concrete definitions by which explanation quality can be measured. They serve as a guide for future research directions for the field. The four principles support the foundation of explainable AI measurement, policy considerations, safety, acceptance by society, and other aspects of AI technology.

We present and discuss the principles in Section 2. We adopt an expansive view of explanations and characterize the space in Section 3. We outline risks introduced by explainable AI – especially those introduced if the principles are not met (Section 4). To put current work into context, we provide a review of current explainable AI methods and evaluation metrics, and other existing principles for explainable AI (Sections 5, 6, and 7). Finally, we review existing literature to assess the extent to which humans meet the same principles we introduce for AI (Section 8). Performance expectations can vary for humans and machines. Although in some contexts, these differing expectations may or may not be appropriate, a baseline on which they could be compared is needed.

2. Four Principles of Explainable AI

We present four fundamental principles for explainable AI systems. These principles are heavily influenced by considering the AI system's interaction with the human recipient of the information. The requirements of the given situation, the task at hand, and the consumer will all influence the type of explanation deemed appropriate for the situation. These situations can include, but are not limited to, regulator and legal requirements, quality control of an AI system, and customer relations. Our four principles are intended to capture a broad set of motivations, reasons, and perspectives. Our principles apply to systems that produce explanations, and they support the full range of AI techniques, not only machine learning ones.

Before we delve into the principles, for this document, we operationally define three key terms: explanation, output, and process. An *explanation* is the evidence, support, or reasoning related to a system's output or process. We define the *output* of a system as i) the outcome from or ii) the action taken by a machine or system performing a task. The output of a system differs by task. For a loan application, the output is a decision: approved or denied. For a recommendation system, the output could be a list of recommended movies. For a grammar checking system, the output is grammatical errors and recommended corrections. For a classification system, it could be an object identifier or a spam detector. For automated driving, it could be the navigation itself. The *process* refers to the procedures, design, and system workflow which underlie the system (c.f. [50]). This includes documentation about the system, information on data used for system development or data stored, and related knowledge about the system.

Briefly, our four principles of explainable AI are:

Explanation: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.

Meaningful: A system provides explanations that are understandable to the intended consumer(s).

Explanation Accuracy: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.

Knowledge Limits: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output.

These are defined and put into context in more detail below. Figure 1 shows the principles and indicates that for a system to be considered explainable, it must first have an explanation or contain accompanying evidence which can be accessed.

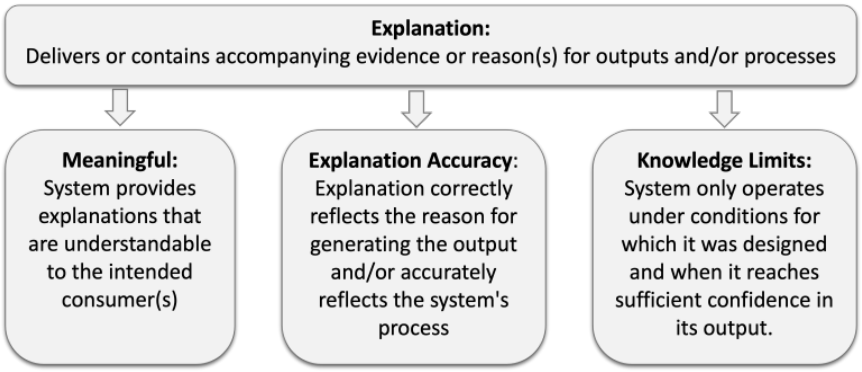

Fig. 1. Illustration of the four principles of explainable artificial intelligence. Arrows indicate that for a system to be explainable, it must provide an explanation. The remaining three principles are the fundamental properties of those explanations.

2.1 Explanation

The *Explanation* principle states that for a system to be considered explainable it supplies evidence, support, or reasoning related to an outcome from or a process of an AI system. By itself, the explanation principle is independent of whether the explanation is correct, informative, or intelligible. This principle does not impose any metric of quality on those explanations. These factors are components of the meaningful and explanation accuracy principles. Explanations in practice will vary, and should, according to the given system and scenario. This means there will be a large range of ways an explanation can be executed or embedded into a system. To accommodate a large range of applications we adopt a deliberately broad definition of an explanation.

2.2 Meaningful

A system fulfills the *Meaningful* principle if the intended recipient understands the system's explanation(s). There are commonalities across explanations which can make them more

meaningful [84]. For example, stating why the system *did* behave a certain way can be more understandable than describing why it *did not* behave a certain way [76]. Many factors contribute to what individual people will consider a "good" explanation [55, 84, 139]. Therefore, developers need to consider the intended audience [44]. Several factors influence what information people will find important, relevant, or useful. These include a person's prior knowledge and experiences and the overall psychological differences between people [18, 64, 90]. Moreover, what they consider meaningful will change over time as they gain experience with a task or system [18]. Different *groups* of people will also have different desires from a system's explanations [13, 44, 50]. Groups may be defined broadly according to their role or relationship to the system. For example: developers of a system are likely to have different desires from an explanation compared to an end-user.

In addition to its audience, what is considered meaningful will vary according to the explanation's purpose. Different scenarios and needs will drive what is important and useful in a given context. Meeting the Meaningful principle will be accomplished by understanding the audience's needs, level of expertise, and relevancy to the question or query at hand. We provide a more detailed discussion of these purposes in Section 3.

Measuring the meaningful principle is an area of ongoing work (Section 7.1). The challenge is to develop measurement protocols that adapt to different audiences. Rather than viewing this as a burden, we argue that both the awareness and appreciation of an explanation's context support the ability to measure the quality of AI explanations. Scoping these factors will therefore bound the possibilities for how to execute the explanation in a goal-oriented and meaningful way.

2.3 Explanation Accuracy

Together, the Explanation and Meaningful principles only call for a system to produce explanations that are intelligible to the intended audience. These two principles do not require that an explanation correctly reflects a system's process for generating its output. The *Explanation Accuracy* principle imposes veracity on a system's explanations.

Explanation accuracy is a distinct concept from decision accuracy. Decision accuracy refers to whether the system's judgment is correct or incorrect. Regardless of the system's decision accuracy, the corresponding explanation may or may not accurately describe *how* the system came to its conclusion or action. Researchers in AI have developed standard measures of algorithm and system accuracy [23, 29, 52, 96, 97, 99, 103, 114]. While these established decision accuracy metrics exist, researchers are in the process of developing performance metrics for explanation accuracy. In Section 7.2, we review current work on this subject.

Additionally, explanation accuracy needs to account for the level of detail in the explanation. For some audiences and/or purposes, simple explanations will suffice. The given reasoning might succinctly focus on the critical point(s) or provide a high level reasoning without extensive detail. These simple explanations could lack nuances that are necessary to completely characterize the algorithm's process for generating its output. However, these

nuances may only be meaningful to certain audiences, such as experts of the system. This is similar to how humans approach explaining complex topics. A professor of neuroscience may explain a new finding with extensive and technical details to a colleague. That same finding will likely be distilled and changed for presenting to an undergraduate student in order to present the pertinent and higher level details. That same professor may explain the finding very differently to their untrained friends and parents.

Together, this highlights the point that explanation accuracy and meaningfulness interact. A detailed explanation may accurately reflect the system's processing, but sacrifice how useful and accessible it is to certain audiences. Likewise, a brief, simple explanation may be highly understandable but would not fully characterize the system. Given these considerations, this principle allows for flexibility in explanation accuracy metrics.

2.4 Knowledge Limits

The previous principles implicitly assume that a system is operating within the scope of its design and knowledge boundaries. The *Knowledge Limits* principle states that systems identify cases in which they were not designed or approved to operate, or in cases for which their answers are not reliable. By identifying and declaring knowledge limits, this practice safeguards answers so that a judgment is not provided when it may be inappropriate to do so. This principle can increase trust in a system by preventing misleading, dangerous, or unjust outputs.

There are two ways a system can reach or exceed its knowledge limits. In one way, the operation or query to the system can be outside its domain. For example, in a system built to classify bird species, a user may input an image of an apple. The system could return an answer to indicate that it could not find any birds in the input image; therefore, the system cannot provide an answer. This is both an answer and an explanation. In a second way, the confidence of the most likely answer may be too low, depending on an internal confidence threshold. To revisit an example of the bird classification system, the input image of a bird may be too blurry to determine its species. In this case, the system may recognize that the image is of a bird but that the image is of low quality. An example output may be: "I found a bird in the image, but the image quality is too low to identify it."

2.5 Summary

Given the wide range of needs and applications of explainable AI systems, a system may be considered more explainable, or better able to meet the principles, if it can generate more than one type of explanation. Further, the metrics used to evaluate the accuracy of an explanation may not be universal or absolute. A body of ongoing work currently seeks to develop and validate explainable AI methods. An overview of these efforts is provided in Sections 6 and 7. The four principles serve as a guidance for how to consider whether the explanation itself meets user needs.

The field of explainable AI is an area of active research. Our understanding of these systems and their use will vary as the field grows with new knowledge and data. Therefore,

these principles serve as a way to guide how we think about the needs of the system. These principles provide a basis for approaching new challenges and questions.

3. Purposes and styles of explanations

To illustrate the broad range of explanations, we characterize explanations by two properties: purpose and style. *Purpose* is the reason why a person requests an explanation or what question the explanation intends to answer. *Style* describes how an explanation is delivered.

The audience will strongly influence the purpose of the explanation and the information it provides. This information will vary according to different groups of people and their role in the system [13, 44, 50]. A system builder may want explanations related to debugging AI models or evaluating training data. Regulators may inquire if a system meets stated regulatory requirements [44].

The explanation's purpose will in turn influence its style. In Figure 2, we visualize our three elements of style: level of detail, degree of interaction between the human and machine, and its format. These attributes are not exhaustive – explanations can take many forms. However, we highlight these elements as ones closely related to meeting the four principles. Therefore, considering these will lay the groundwork for producing explanations. We expound upon these in more detail below.

The *level of detail* is depicted as a range, from sparse to extensive. By sparse, we mean that the amount of information provided is brief, limited, and/or at a high-level, lacking in detail. An example of a sparse explanation might be an explanation for a decision made by an alert system (e.g., "system processes slowed because of overheating."). An extensive explanation may contain detailed information about a system and/or provide a large amount of information (e.g., a report with relevant system information to understand its process).

We place the *degree of human-machine interaction* into three categories: declarative explanations, one-way interaction, and two-way interaction. In a declarative explanation, the systems provides an explanation, and there is no further interaction. This describes most current explainable AI methods (Section 6). For example, a loan application system may always output the rationale for an acceptance or rejection. An object classifier may output a saliency map [120]. A model card [86] may contain pre-determined information about the system. A declarative explanation is based on a default query, such as "why did the object classifier produce this decision?". The human cannot alter the question being asked (barring a change in the system itself to produce something different).

A higher degree of interaction is a *one-way interaction*. For this, the explanation is determined based on a query or question input to the system [20, 35, 142]. For example, this may be a graphical output depending on the factors a person wishes to visualize. This may allow the explanation's consumer the ability to probe further or to submit different queries.

We define the category with the deepest interaction level as the *two-way interaction*. This models a conversation between people. The person can probe further, and the machine can probe back, ask clarifying questions, or provide new avenues of exploration. For

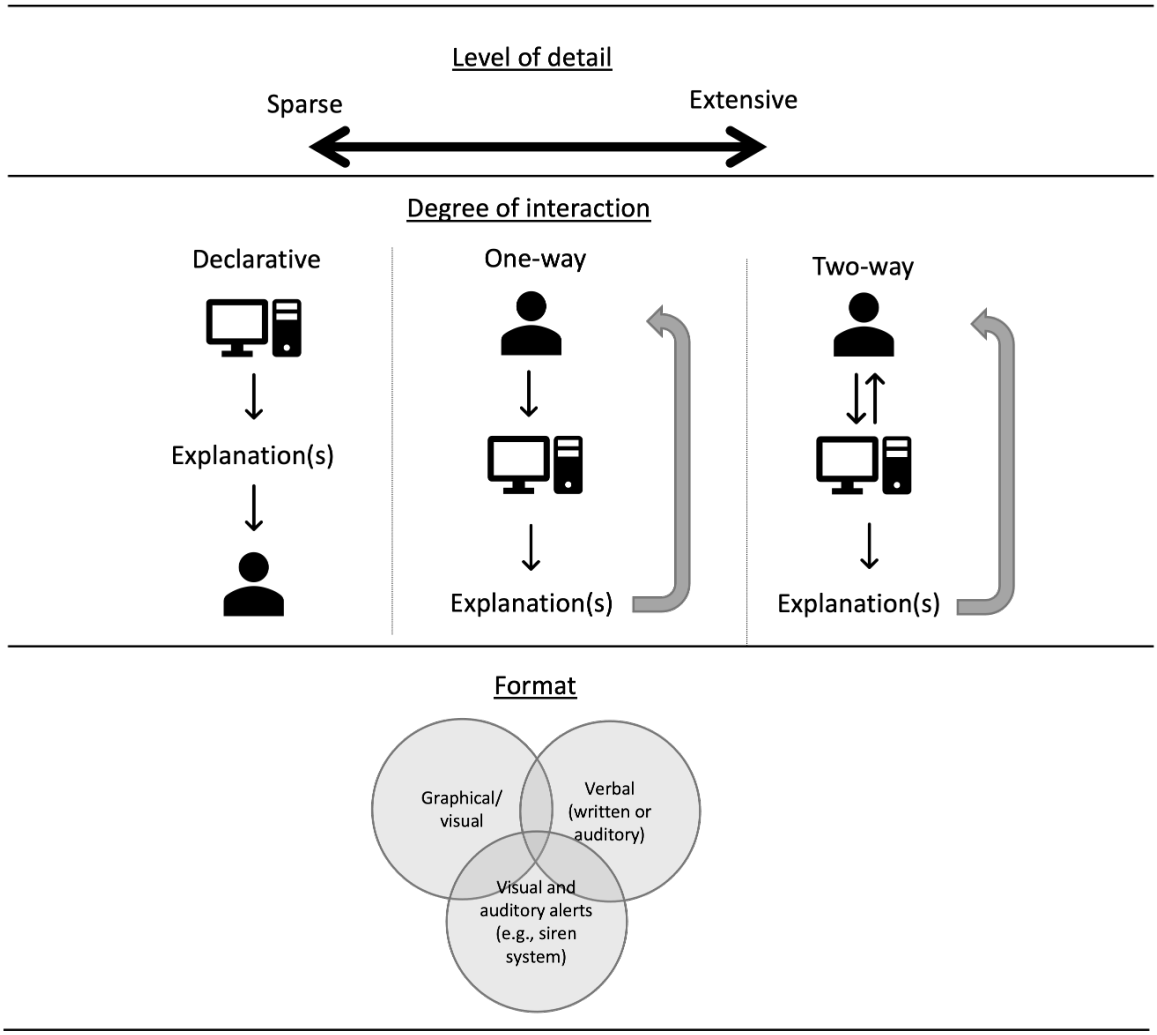

Fig. 2. Illustration of our elements of explanation styles.

example, a system may probe the user for additional details or propose alternate questions. To our knowledge, two-way interactions do not yet exist. Developing them is a future research direction.

The explanation's *format* includes visual and graphical, verbal, and auditory or visual alerts. Examples of graphical formats include outputs from data analyses or saliency maps. Verbal formats can include written outputs and reports as well as auditory outputs, such as speech. These visual and verbal formats carry the assumption that the audience is expecting and attending to the explanation. Another form of explanation can capture an unaware audience's attention. A siren or light system can produce different alarms, light flashing patterns, and/or light colors as an explanation that alert the audience. For example, a specific siren pitch or pattern could indicate something about a system's status which may need attention.

Each of these style elements will need to be considered to produce an explanation for its purpose and to meet the four principles. Some cases may call for a simple, declarative explanation as the most appropriate style to optimize how meaningful it is. This is sometimes the case in a weather emergency, such as when a tornado is in the area. A current weather alert from the National Weather Service, "Tornado Warning: Take Action!"[3], operates as both an alert and a simple explanation. The alert is to "Take Action" with the simple explanation of "Tornado Warning." Depending on the metric, this explanation may not be considered highly accurate because the minimal level of detail: it does not include why a tornado warning is declared in its explanation. However, in this example, brevity is appropriate to ensure it is understood by a wide audience and to enable swift action. Although minimal, additional information with an alert may be helpful to address non-compliance in responding to weather alerts (e.g., "cry-wolf" effects) [73]. In a different scenario, such as when debugging a system, the explanation could include information on the internal steps of a system. This could be lengthy and contain field-specific language. The audience may need time and more effort to examine the explanation and to decide on their next actions. Here, more details in the user's preferred format would be helpful, and two-way interactions could become important.

These different purposes, styles, and considerations illustrate the range and types of explanations. This points to the need for flexibility in addressing the scope of systems that require explanations. Because the circumstances under which an explanation is provided will differ, the four principles encourage adapting to different styles as appropriate. Some explanations will be easier to achieve than others, and designers will need to consider trade-offs between accomplishing different goals.

4. Risk Management of Explainable AI

Risk is defined as "the effect of uncertainty of objectives" [22, pg. 6] and includes both negative outcomes (threats) as well as positive outcomes (opportunities). Risk management is a process that can be used to define, assess, and mitigate risk. Explainable AI can mitigate

[3] https://www.weather.gov/safety/tornado-ww

the risks of artificial intelligence by assessing, measuring, or predicting the risk in a model or system. Explanations can be used to test for vulnerabilities [41]. Alternatively, explainable AI can introduce risks of its own, *e.g.*, adversarial attacks discussed in Section 6.3. This section focuses on the latter, managing the potential risks *introduced by* explainable AI.

Any explainable AI system will contain potential risks, both threats and opportunities. The degree to which stakeholders are prepared to accept the trade-off of general risk and goals is called the stakeholders' *risk appetite* [22]. Many risk management strategies share the common components: identifying, analyzing, responding, monitoring, and reviewing. For explainable AI, a risk management strategy will need to factor the four principles.

An *explanation*, the first principle, is necessary for explainable AI, but an explanation itself introduces risks, both positive and negative. A potential negative outcome of having an explanation is the exposure of proprietary details [85]. A single explanation may not expose the inner workings of the system. However, multiple explanations, either from multiple independent queries or through a two-way interaction, could expose intellectual property when connected to each other. How many explanations must an end user have access to before they have an understanding of the system? The scope of each explanation may impact the number. These include explanations that describe the limit of the system's knowledge.

However, explanations also have the potential for positive outcomes. A user can better understand the system. This could lead to improvements, such as increased trust in the system. Explanations may also be necessary for compliance with regulations, *e.g.*, The Fair Credit Reporting Act (FCRA) and Article 13 of the European Union General Data Protection Regulation (GDPR).

Explanations need to be *meaningful* for the audience. This is the second principle and introduces its own risks. A meaningful explanation can give deeper insight into the system, but it may expose intellectual property or system vulnerabilities by exposing its inner workings. An explanation that is not meaningful, on the other hand, is in jeopardy of being ignored or not recognized as an explanation.

In order to be useful, an explanation not only needs to be meaningful, it must also be *accurate*, our third principle. A relevant potential risk is commonly known as *model risk*, the potential negative outcomes derived from an invalid or misapplied model. As stated in the Federal Reserve System [31], one of the two main sources for this type of risk is underlying errors in the model causing erroneous outputs. An inaccurate explanation can lead to a misinterpretation or misunderstanding of how the system works or arrived at an outcome. This is a negative risk for an end user, but AI systems are also used as one part of a larger system where the AI might bias a human.

In face recognition, a human face examiner could receive information from an AI algorithm on which parts of the face are useful. An accurate explanation can help the examiner more accurately assess the pair of faces while an inaccurate explanation could lead to a wrong decision. In the judicial system, AI algorithms have been used in decision, such as if a defendant may be arrested again [8]. An inaccurate explanation of how the algorithm

arrived at its outcome could result in a miscarriage of justice. An accurate explanation can help create a more just society.

The other main source of model risk is using the model incorrectly or beyond its knowledge limits [31]. Explanations that describe the *knowledge limits* of the system, the fourth principle, can provide assurance the model is not operating out of scope and nurture confidence. Describing the limits of a system can potentially expose the inner workings of the system, especially if combined with information collected from other explanations.

Examining the potential risks of software exposure, there are different contexts, categories, and levels of risk. Who are the end users? If they are only internal to the organizations such as developers, the management strategy will be different than if the end users include external customers.

Explainable AI introduces new threats to a system. However, it also introduces new opportunities as well. Whether the outcome is a threat or an opportunity sometimes depends on the audience. Risk management considers the trade-offs and possibilities of these and other factors. When assessing risk, two components often assessed are the likelihood of the risk and the impact of the outcome [37].

In general for AI, there is a need to develop a risk management framework; a request for information by NIST[4] occurred on 2021-07-29. For more information on risk management, see [22, 37, 128].

5. Overview of Principles in the Literature

Theories and properties of explainable AI have been discussed from different perspectives, with commonalities and differences across these points of view [9, 26, 34, 50, 79, 111, 112, 141].

Lipton [79] divides explainable techniques into two broad categories: transparent and post-hoc interpretability. Lipton [79] defines a transparent explanation as reflecting to some degree how a system came to its output. A subclass is simulatability, which requires that a person can grasp the entire model. This implies that explanations will reflect the inner workings of a system. Their post-hoc explanations "often do not elucidate precisely how a model works, they may nonetheless confer useful information for practitioners and end users of machine learning." For example, the bird is a cardinal because it is similar to cardinals in the training set.

Rudin [111] argues that it should not be assumed that interpretability must be sacrificed for state-of-the-art accuracy. They recommend that for high stakes decisions, one should avoid a black-box model, unless one can prove that an interpretable model does not exist with the same level of accuracy. Note that we will we refer to black-box as *closed-box* for the remainder of this document [94]. Rudin et al. [113] builds on their previous work by presenting five principles and ten grand challenges of interpretable machine learning.

Mueller et al. [91] reviews some of the basic concepts for user-centered explainable AI systems. Based on these concepts, they describe the Self-Explanation Scorecard, and

[4] https://www.federalregister.gov/d/2021-16176; Date Accessed: 2021-08-31

present a set of user-centered design principles.

From a psychological perspective, Broniatowski [17] makes the case that interpretability and explainability are distinct requirements for machine learning systems. The resulting analysis implies that system output should be tailored to different types of users.

Wachter et al. [140] argue that explanations do need to meet the explanation accuracy property. They claim that counterfactual explanations are sufficient. "A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the prediction to a predefined output [88];" e.g., if you had arrived to the platform 15 minutes earlier, you would have caught the train. Counterfactual explanations do not necessarily reveal the inner workings of a system. This property allows counterfactual explanations to protect intellectual property.

Gilpin et al. [34] defines a set of concepts for explainable AI. Similar to the meaningful and explanation accuracy principles in our current work, Gilpin et al. [34] propose that explanations should allow for a trade-off between their interpretability and completeness. In addition, they state that trade-offs must not obscure key limitations of a system.

Doshi-Velez and Kim [26] address the critical question: measuring if explanations are meaningful for users or consumers. They present a framework for a science to measure the efficiency of explanations. This paper discusses factors that are required to begin testing interpretability of explainable systems. The paper highlights the gap between these principles as a concept and creating metrics and evaluation methods.

Information Commissioner's Office and The Alan Turing Institute [50] lays out principles to follow for explainable AI. These principles are: be transparent, be accountable, consider the context you are operating in, and reflect on the impact of your AI system on the individuals affected as well as the wider society [50]. In addition to discussing principles, they discuss different things that go into an explanation, including process-based explanations vs. outcome-based explanations, the rationale, the responsibility of who made what decisions, an explanation of the data, and design steps that maximize fairness, safety, and impact of the use of the system.

Barredo Arrieta et al. [9] discuss these various terms used in different sources to describe explainability or interpretability: understandability, comprehensibility, interpretability, explainability, and transparency. They discuss how these terms are all different yet tied together.

Weller [141] discusses types of transparency and how they address different classes of users or consumers of explanations. Similar to the explanation accuracy principle, the paper introduces faithfulness of an explanation as

> ...broadly beneficial for society provided that explanations given are *faithful*, in the sense that they accurately convey a true understanding without hiding important details. This notion of faithful can be hard to characterize precisely. It is similar in spirit to the instructions sometimes given in courts to tell "the truth, the whole truth, and nothing but the truth." [141]

Across these viewpoints, there exist both commonalities and disagreement. Similar to

our four principles, commonalities include concepts which distinguish between the existence of an explanation, how meaningful it is, and how accurate or complete it is. Although disagreements remain, these perspectives provide guidance for development of explainable systems. A key disagreement between philosophies is the relative importance of explanation meaningfulness and accuracy. These disagreements highlight the difficulty in balancing multiple principles simultaneously. Context of the application, community and user requirements, and the specific task will drive the importance of each principle.

6. Overview of Explainable AI Algorithms

Researchers have developed different algorithms to explain AI systems. Following other sources [9, 38, 68], we organize the explanations into two broad categories: self-interpretable models and post-hoc explanations. *Self-interpretable models* are the algorithm model (or a representation of the algorithm itself) that can be directly read and interpreted by a human. In this case the model itself is the explanation. *Post-hoc explanations* are explanations, often generated by other software tools, that describe, explain, or model the algorithm to give an idea of how the algorithm works. Post-hoc explanations often can be used on algorithms without any inner knowledge of how the algorithm works, provided that it can be queried for outputs on chosen inputs.

Rather than mention all of the different explanation subtypes and all of the different explanations available, we highlight a few widely-used examples, some categorizations, and then refer the reader to various surveys on explainable AI [2, 9, 38, 68, 78, 89].

6.1 Self-Interpretable Models

Self-interpretable models are models that are themselves the explanations. Not only do they explain the entire model globally, but by walking through each input through the model, the simulation of the input on the self-interpretable model can provide a local explanation for each decision.

Some of the most common self-interpretable models include decision trees and regression models (including logistic regression). There is work in producing many more interpretable models that improve in accuracy over basic decision trees and basic regression models in many cases. These models include decision lists [70], decision sets [71], prototypes (representative samples of each class) Kim et al. [56], feature combination rules that completely classify sets of inputs Kuhn et al. [61], Bayesian Rule Lists [74], additive decision trees [81] and improved variants of decision trees [4, 11, 77].

With self-interpretable models, some sources state an *accuracy-interpretability trade-off* [19, 27, 78]: self-interpretable models are less accurate than post-hoc models because there is a trade-off between making the model more exact or more meaningful to humans. However, Rudin [111], Rudin and Radin [112] disagree, arguing that there is not necessarily an accuracy-interpretability trade-off and in many cases interpretable models can be used without a loss of decision accuracy.

6.2 Post-Hoc Explanations

Post-hoc explanations are grouped into two kinds: local explanations and global explanations. A *local explanation* explains a subset of decisions or is a per-decision explanation. A *global explanation* produces a model that approximates the non-interpretable model. In some cases, a global explanation can also provide local explanations by simulating them on specific inputs to provide local explanations for those individual inputs. As simple examples, consider a logistic regression (which could either be a self-interpretable model or a post-hoc approximation to an opaque model). The regression coefficients provide a global explanation that explain all inputs. However, one can plug the input in with the weights and then use those weights to explain the output of the algorithm.

We discuss each of these explanations in the following subsections, describing local explanations in Section 6.2.1 and global explanations in Section 6.2.2.

6.2.1 Local Explanations

Local explanations explain a subset of inputs. The most common type of local explanation is a *per-decision* or *single-decision* explanation, which provides an explanation for the algorithm output or decision on a single input point.

One commonly-used local explanation algorithm is LIME (Local Interpretable Model-Agnostic Explainer) [107]. LIME takes a decision, and by querying nearby points, builds an interpretable model that represents the local decision, and then uses that model to provide per-feature explanations. The default model chosen is logistic regression. For images, LIME breaks each image into superpixels, and then queries the model with a random search space where it varies which superpixels are omitted and replaced with all black (or a color of the user's choice).

Another commonly-used local explanation algorithm is SHAP (SHapley Additive exPlanations) [82]. SHAP provides a per-feature importance for an input on a regression problem by converting the scenario to a coalitional game from game theory and then producing the Shapley values from that game. SHAP treats the features as the players, the features value vs. a default value as the strategies, and the system output as the payoff, forming a coalitional game from the input. See Ferguson [32] for more information on Shapley values and coalitional games.

Another common local explanation is a counterfactual. A counterfactual is an explanation saying "if the input were this new input instead, the system would have made a different decision." [140] In these explanations, although there are often multiple widely-differing instances that all are counterfactuals, a counterfactual explanation often provides a single instance. The hope is that the instance is as similar as possible to the input with the exception that the system makes a different decision. However, some systems can produce multiple counterfactual instances as a single explanation. Ustun et al. [138] develop a counterfactual explanation of logistic (or linear) regression models. Counterfactuals are represented as the amounts of specific features to change.

Another popular type of local explanations for problems on image data are *saliency*

pixels. Saliency pixels color each pixel depending on how much that pixel contributes to the classification decision. One of the first saliency algorithms is Class Activation Maps (CAM) [150]. A popular saliency pixel algorithm that enhanced CAM is GRAD-CAM [120]. GRAD-CAM generalized CAM so that it can explain any convolutional network.

An additional local explanation in Koh and Liang [58] takes a decision and produces an estimate of the influence of each training data point on that particular decision. Another additional local explanation is an Individual Conditional Expectation (ICE) [89, 149]. An ICE curve shows the marginal effect of the change in one feature for an instance of the data.

6.2.2 Global Explanations

Global explanations produce post-hoc explanations on the entire algorithm. Often, this involves producing a global model for an algorithm or a system.

One Global explanation is Partial Dependence Plots (PDPs) [89, 149]. A Partial Dependence Plots shows the marginal change of the predicted response when the feature (value of that specific data column or component) changes. PDPs are useful for determining if a relationship between a feature and the response is linear or more complex [89].

In deep neural networks, one such global algorithm is TCAV (Testing with Concept Activation Vectors) [153]. TCAV wishes to explain a neural network in a more user-friendly way by representing the neural network state as a linear weighting of human-friendly concepts, called Concept Activation Vectors (CAVs). TCAV was applied to explain image classification algorithms through learning CAVs including color, to see how colors influenced the image classifier's decisions.

Two visualizations used to provide global explanations are Partial Dependence Plots (PDPs) and Individual Conditional Expectation (ICE) [89, 149]. The partial dependence plot shows the marginal change of the predicted response when the feature (value of that specific data column or component) changes. PDPs are useful for determining if a relationship between a feature and the response is linear or more complex [89]. The ICE curves are finer-grained and show the marginal effect of the change in one feature for each instance of the data. ICE curves are useful to check if the relationship visualized in the PCP is the same across all ICE curves, and can help identify potential interactions.

Prototypes [75], representative samples of each class, are also sometimes used as a global explanation for a neural network in addition to sometimes being a self-interpretable model as mentioned in Section 6.1.

Another way to produce global explanations is to summarize local explanations taken on a variety of inputs. A variant of LIME, SP-LIME [107], uses a submodular pick to choose the most relevant local LIME explanations as summary explanations. Another way is to try to approximate the post-hoc model by learning a global model on a system such as a decision set [72] or a summary of counterfactual rules [106].

6.3 Adversarial Attacks on Explainability

Explanation accuracy (Principle 3) is an important component of explanations. Sometimes, if an explanation does not have 100 percent explanation accuracy, it can be exploited by adversaries who manipulate a classifier's output on small perturbations of an input to hide the biases of a system. First, Lakkaraju and Bastani [69] observes that even if an explanation can mimic the predictions of the closed-box that this is insufficient for explanation accuracy and such systems can produce explanations that may mislead users. An approach to generate misleading explanations is demonstrated in Slack et al. [124]. They do this by producing a scaffolding around a given classifier that matches the classification on all input data instances but changes outputs for small perturbations of input points, which can obfuscate global system behavior when only queried locally. This means that if the system is anticipating being explained by a tool such as LIME that gives similar instances to training set instances as inputs, the system will develop an alternative protocol to decide those instances that differ from how they will classify trials in the training and test sets. This can mislead the explainer by anticipating which trials the system might be asked to classify. Another similar approach is demonstrated in Aivodji et al. [5]. They fairwash a model by taking a closed-box model and produce an ensemble of interpretable models that approximate the original model but are much fairer, which then hide the unfairness of the original model. Another example of slightly perturbing a model to manipulate explanations is demonstrated in Dimanov et al. [24]. The ability for developers to cover up unfairness in closed-box models is one of the several vulnerabilities of explainable AI discussed in Hall et al. [41]. Kindermans et al. [57] shows that many saliency pixel explanations lack input invariance, meaning that a small change to the input can greatly change the output and the attribution to relevant pixels.

7. Evaluating Explainable AI Algorithms

This sections summarizes the state-of-the art of evaluating explainable AI algorithms (cf., [151]). In this paper, we separate the evaluation of explainable AI algorithms according to which principle is being evaluated. The Explanation principle (Principle 1) is covered under the section, Overview of Explainable AI Algorithms (Section 6), which reviews current explanation methods. In this section, we review current methods for measuring explanation meaningfulness (Principle 2) and explanation accuracy (Principle 3). To our knowledge there is limited work on developing and evaluating algorithms' knowledge limits (Principle 4). As a result, we do not discuss evaluating knowledge limits in this section.

7.1 Evaluating Meaningfulness

One way to measure the meaningfulness of an explanation involves measuring *human simulatability*. This is essentially the ability for a person to understand a machine learning model to the extent that they would be able to take the same input data as the model, and understand the parameters of the model such that they would be able to produce a prediction

from the model themselves in a reasonable amount of time [79]. The ability to simulate the model themselves would reflect a high degree of understanding. This is typically measured for self-interpretable models as a way to measure the complexity of the model.

Several studies have put human simulatability to the test. Lage et al. [65] and Lage et al. [66] measured the accuracy of the humans' results, the response time taken, and a human poll of the subjective difficulty of simulating the model. Hase and Bansal [43] discusses two kinds of human simulatability: *forward simulation*, which is when a human predicts a system's output for a given input; and *counterfactual simulation*, where a human is given an input and an output. They must predict what output the system would give if the input were changed in a particular way. When evaluating explanations, they evaluated forward and counterfactual simulation by measuring the change in user accuracy relative to different explanations. Poursabzi-Sangdeh et al. [101] measured the accuracy of humans simulating different logistic regression models on housing prices. Slack et al. [123] conducted a "what-if" simulatability evaluation: the user receives an input with an explanation. The user is then asked to simulate the model on a new input that is slightly perturbed from the given input (the new input mirrors a what-if or counterfactual).

Another strategy to evaluate meaningfulness is to ask humans to complete a task using the provided system's output as input, then measuring the human's time taken and decision accuracy on the task. Poursabzi-Sangdeh et al. [101] does this by also asking humans to predict what they believe house prices should be, in addition to asking what the model will predict the house price will be (humans can disagree with the models in this step). Kim et al. [56] harnessed the power of examples. Their model, the Bayesian Case Model (BCM), learned prototypes of different cooking recipes. Humans were provided only the ingredients of the prototype and were measured on how well they were able to classify each recipe. Lai and Tan [67] tested meaningfulness in a deception detection task. The task was to determine if hotel reviews were genuine or deceptive. Human accuracy of deception detection was compared when they were only provided the review itself and when they were presented with explanations from a machine. This comparison enables comparing human decision accuracy with and without machine assistance/explanations. Lakkaraju et al. [72] evaluated the interpretability of different complexity decision sets by asking humans to view explanations and make decisions, measuring their accuracy and response time. Mac Aodha et al. [83] evaluated an explanation by comparing human accuracy when humans are trained with systems that provide explanations compared to being trained with systems that do not provide an explanation. Schmidt and Biessmann [116] recruited users to complete tasks given with and without system explanations and measures each user's total time taken and decision accuracy. Anderson et al. [7] studied two techniques for explaining the actions of reinforcement learning agents to people not trained in AI. They tested multiple explanation conditions: no explanation, each of the two explanations separately, and both explanations. Overall, humans were most accurate when combining both techniques, saliency maps and reward-decomposition bars.

Meaningfulness has been measured with subjective ratings as well. Hoffman et al. [45] discussed a variety of criteria for good explanations and provide an Explanation Satisfac-

tion Scale. Holzinger et al. [46] developed the System Causability Scale (SCS) to compare explanations. As part of their evaluation of human simulatability, Lage et al. [65] also asked humans to subjectively rate the difficulty of simulating a model. Rajagopal et al. [105] conducted experiments asking users to evaluate different properties of explanations.

Metrics on the size or complexity of a model are sometimes used as measures for a model's interpretability. Lakkaraju et al. [71] measured the interpretability of a model by asking users if the information provided was sufficient to make conclusions. Poursabzi-Sangdeh et al. [101] compared two types of models to test which enabled participants more closely simulate the model's actual predictions. They found that less information (a less transparent model) could enable this better than a more transparent model perhaps due to "information overload"). Lage et al. [66] measured the effect of complexity on human simulatability. The idea is that different levels, and types, of complexity can affect transparency more or less than other types. Lakkaraju et al. [72] asked humans to made decisions, provided them the explanation as help, and measured how quickly and how accurately they made decisions. Narayanan et al. [92] compared different types of output complexity for how they affected human performance. Bhatt et al. [12] designed a complexity metric to quantify "feature importance" explanations.

7.2 Evaluating Explanation Accuracy

Explanation accuracy is closely related to work on "fidelity". Several studies have evaluated explanation fidelity [87, 110]. One way this has been tested is to simulate models by using the system output as the ground truth and evaluating the post-hoc explanations using a machine learning metric [87, 107]. Lakkaraju et al. [72] followed this strategy but also checked that each instance had at most one explanation and that every instance is explained by the post-hoc explanation model. The second method Mohseni et al. [87] proposed is having humans evaluate the explanations and apply "sanity checks" to evaluate the explanation accuracy. The third method asked the system to explain a variety of inputs. In many cases, the inputs are adaptive. New inputs are slightly changed versions of the previous inputs, based on the provided explanation. Experiments then measure the change in output relative to the change in input and the importance of the changed features from the explanation. Samek et al. [115] evaluated the quality of the explanation accuracy with saliency pixels. They gradually deleted the most important pixels and measured how much the classification score changes. The idea is that if pixels which are important have more influence on decision accuracy, and as they are deleted, the system is less likely to classify the image as the original class. Hooker et al. [47] tested whether systems performed worse when important features are removed. They applied a strategy in which they removed important pixels, then retrained systems and measured decision accuracy of the retrained systems. Yeh et al. [147] developed an "infidelity measure" to evaluate explanation accuracy. Alvarez Melis and Jaakkola [6] evaluated the explanation accuracy, or faithfulness, by removing the model's higher order features and measuring the drop in classification probability. They also measured explanation accuracy by adding white noise to

the inputs and measuring how much the explanation changes. Adebayo et al. [3] evaluated explanation accuracy of saliency pixel explanations for deep neural networks by measuring the amount the explanation changed relative to how the trained models differed. Sixt et al. [122] evaluated the quality of saliency pixels by randomizing middle convolutional layers and comparing saliency pixels. They also compared the saliency pixels when the labels are the actual labels vs. random. Qi et al. [104] evaluated explanation accuracy by adding or deleting image pixels deemed relevant by the explanation. They then compared the system's scores on the new images. Bhatt et al. [12] evaluated the explanation accuracy of "feature importance" explanations by both checking sensitivity, meaning similar inputs have similar feature importance explanations, and faithfulness, meaning the change in the explanations should correlate to the change in inputs.

The quality of counterfactual explanations were tested in Wachter et al. [140]. A counterfactual explanation should answer, "what is the minimum amount an input would need to change for the system to change its decision on that input?" Therefore, they tested how far away the counterfactual was from the original data point.

8. Humans as a Comparison Group for Explainable AI

When considering the performance of humans and AI systems, there are fairly significant differences of opinion regarding performance expectations. Some argue that we should hold machines to a much higher standard than humans, while others believe it is sufficient for machines to simply be as good as humans. A cascade of interesting and difficult questions arise from this overarching philosophical divide, such as how much better do machines have to be than humans? In what way(s) must they be better? How do we measure "as good as"? Regardless of where one falls on this particular philosophical debate, it is nonetheless helpful to consider human performance as a baseline. In this section, we describe human decision-making with respect to the extent humans explanations line up with our four principles.

Independent of AI, humans operating alone also make high stakes decisions with the expectation that they be explainable. For example, physicians, judges, lawyers, and forensic scientists are often expected to provide a rationale for their judgments. How do these proffered explanations adhere to our four principles? We focused strictly on human explanations of their own judgments and decisions (e.g.,"why did you arrive at this conclusion or choice?"), not of external events (e.g., "why is the sky blue?" or "why did an event occur?"). External events accompanied by explanations can be helpful for human reasoning and formulating predictions [80]. This is consistent with a desire for explainable AI. However, as outlined in what follows, human-produced explanations for their own judgments, decisions, and conclusions are largely unreliable. Humans as a comparison group for explainable AI can inform the development of benchmark metrics for explainable AI systems; and lead to a better understanding of the dynamics of human-machine collaboration.

8.1 Explanation

This principle states only that for a system to be considered explainable, it provides an explanation. In this section, we will focus on whether humans produce explanations of their own judgments and decisions and whether doing so is beneficial for the decision makers themselves. In Section 8.2, we will discuss whether human explanations are meaningful, and in Section 8.3, we will discuss the accuracy of those explanations.

Humans are able to produce a variety of explanation types [55, 79, 84]. However, producing verbal explanations can interfere with decision and reasoning processes [117, 118, 144]. It is thought that as one gains expertise, the underlying processes become more automatic, outside of conscious awareness, and therefore, more difficult to explain verbally [28, 30, 63, 117]. This produces a similar tension which exists for AI itself: the desire for high accuracy are often thought to come with reductions in explainability (however, c.f., [79]).

More generally, processes which occur with limited conscious awareness can be harmed by requiring the decision itself to be expressed explicitly. An example of this comes from lie detection. Lie detection based on explicitly judging whether or not a person is telling the truth or a lie is typically inaccurate [16, 130]. However, when judgments are provided via implicit categorization tasks, therefore bypassing an explicit judgment, lie detection accuracy can be improved [129, 130]. This suggests that lie detection may be a nonconscious process which is interrupted when forced to be made a conscious one.

Together these findings suggest that some assessments from humans may be more accurate when left automatic and implicit, compared to requiring an explicit judgment or explanation. Human judgments and decision making can oftentimes operate as a closed-box [79], and interfering with this closed-box process can be deleterious to the accuracy of a decision.

8.2 Meaningful

To meet this principle, the system provides explanations that are intelligible and understandable to the intended audience. For this, we focused on the ability of humans to interpret how another human arrived at a conclusion. Here, consider this to mean: 1) whether the audience reaches the same conclusion as intended by the person providing the explanation, and 2) whether the audience agrees with each other on what the conclusion is, based on an explanation.

One analogous case to explainable AI for human-to-human interaction is that of a forensic scientist explaining forensic evidence to laypeople (e.g., members of a jury). Currently, there are gaps between the ways forensic scientists report results and the understanding of those results by laypeople (see Edmond et al. [28], Jackson et al. [51] for reviews). Jackson et al. [51] extensively studied the types of evidence presented to juries and the ability for juries to understand that evidence. They found that most types of explanations from forensic scientists are misleading or prone to confusion. Therefore, they do not meet our internal criteria for being "meaningful." A challenge for the field is learning how to improve

explanations, and the proposed solutions do not always have consistent outcomes [51].

Complications for producing meaningful explanations for others include people expecting different explanation types, depending on the question at hand [84], context driving the formation of opinions [51], and individual differences in what is considered to be a satisfactory explanation [90]. Therefore, what is considered meaningful varies by context and across people.

8.3 Explanation Accuracy

This principle states that a system's explanation correctly reflects its reasons for generating a certain output and/or accurately reflects its process. For humans, this is analogous to an explanation of one's decision processes truly reflecting the mental processes behind that decision. In this section, we focused on this aspect only. An evaluation of the quality or coherence of the explanation falls outside of the scope of this principle.

For the type of introspection related to explanation accuracy, it is well-documented that although people often report their reasoning for decisions, this does not reliably reflect accurate or meaningful introspection [93, 102, 143]. This has been coined the "introspection illusion": a term to indicate that information gained by looking inward to one's mental contents is based on mistaken notions that doing so has value [102]. People fabricate reasons for their decisions, even those thought to be immutable, such as personally held opinions [40, 53, 143]. In fact, people's conscious reasoning that is able to be verbalized does not seem to always occur before the expressed decision. Instead, evidence suggests that people make their decision and then apply reasons for those decisions *after* the fact [137]. From a neuroscience perspective, neural markers of a decision can occur up to 10 seconds before a person's conscious awareness [125]. This finding suggests that decision making processes begin long before our conscious awareness.

People are largely unaware of their inability to introspect accurately. This is documented through studies of "choice blindness" in which people do not accurately recall their prior decisions. Despite this inaccurate recollection, participants will provide reasons for making selections they never, in fact, made [39, 40, 53]. For studies that do not involve long-term memory, participants have also been shown to be unaware of the ways they evaluate perceptual judgments. For example, people are inaccurate when reporting which facial features they use to determine someone's identity [108, 135].

Based on our definition of explanation accuracy, these findings do not support the idea that humans reliably meet this criteria. As is the case with algorithms, human decision accuracy and explanation accuracy are distinct. For numerous tasks, humans can be highly accurate but cannot verbalize their decision process.

8.4 Knowledge Limits

This principle states that the system only operates 1) under the conditions it was designed and 2) when it reaches a sufficient confidence in its output or actions. For this principle, we narrowed down the broad field of *metacognition*, or thinking about one's own thinking.

Here, we focused on whether humans correctly assess their own ability and accuracy, and whether they know when to report that they do not know an answer. There are several ways to test whether people can evaluate their own knowledge limits. One method is to ask participants to predict how well they believe they performed or will perform on a task, relative to others (e.g., in what percentile will their scores fall relative to other task-takers). Another way to test the awareness of knowledge limits is to obtain a measure of their response confidence, with higher confidence indicating that a person believes with greater likelihood that they are correct.

As demonstrated by the well known Dunning-Kruger Effect [60], most people inaccurately estimate their own ability relative to others. A similar finding is that people, including experts, generally do not *predict* their own accuracy and ability well when asked to explicitly estimate performance [14, 15, 21, 42, 95]. However, a recent replication of the Dunning-Kruger Effect for face perception showed that, although people did not reliably predict their accuracy, their ability estimates varied accordingly with the task difficulty [152]. This suggests that although the exact value (e.g., predicted performance percentile relative to others, or predicted percent correct) may be erroneous, people can modulate the direction of their predicted performance appropriately (e.g., knowing a task was more or less difficult for them).

For a variety of judgments and decisions, people often know when they have made errors, even in the absence of feedback [148]. To use eyewitness testimony as a relevant example: although confidence and accuracy have repeatedly shown to be weakly related [126], a person's confidence does predict their accuracy in the absence of "contamination" through the interrogation process and extended time between the event and the time of recollection [145]. Therefore, human shortcomings in assessing their knowledge limits are similar to those of producing explanations themselves. When asked explicitly to produce an explanation, these explanations can interfere with more automatic processes gained by expertise; they often do not accurately reflect the true cognitive processes. Likewise, as outlined in this section, when people are asked to explicitly predict or estimate their ability level relative to others, they are often inaccurate. However, when asked to assess their confidence for a given decision vs. this explicit judgment, people can gauge their accuracy at levels above chance. This suggests people do have insight into their own knowledge limits, although this insight can be limited or weak in some cases.

9. Discussion and Conclusions

We introduced four principles to encapsulate the fundamental elements for explainable AI systems. The principles provide a framework with which to address different components of an explainable system. These four principles are that the system produce an explanation, that the explanation be meaningful to humans, that the explanation reflects the system's processes accurately, and that the system expresses its knowledge limits. The principles derive their strength when a system follows all four. A system that provides an explanation but is not understandable, not accurate, or outside knowledge limits has reduced value. In

fact, it may impact users' acceptance of a system's outcomes.

There are different approaches and philosophies for developing and evaluating explainable AI. Computer science approaches tackle the problem of explainable AI from a variety of computational and graphical techniques and perspectives, which may lead to promising breakthroughs. A blossoming field puts humans at the forefront when considering the effectiveness of AI explanations and the human factors behind their effectiveness. Our four principles provide a multidisciplinary framework with which to explore this type of human-machine interaction. The practical needs of the system will influence how these principles are addressed (or dismissed). With these needs in mind, the community will ultimately adapt and apply the four principles to capture a wide scope of applications.

The focus of explainable AI has been to advance the capability of the systems to produce a quality explanation. Here, we addressed whether humans can meet the same set of principles we set forth for AI. We showed that humans demonstrate only limited ability to meet the principles outlined here. This provides a benchmark with which to compare AI systems. In reflection of societal expectations, recent regulations have imposed a degree of accountability on AI systems via the requirement for explainable AI [1]. As advances are made in explainable AI, we may find that certain parts of AI systems are better able to meet societal expectations and goals compared to humans. By understanding the explainability of both the AI system and the human in the human-machine collaboration, this opens the door to pursue implementations which incorporate the strengths of each, potentially improving explainability beyond the capability of either the human or AI system in isolation.

In this paper, we focused on a limited set of human factors related to explainable decisions. Much is to be learned and studied regarding the interaction between humans and explainable machines. Although beyond the scope of the current paper, in considering the interface between AI and humans, understanding general principles that drive human reasoning and decision making may prove to be highly informative for the field of explainable AI [36]. For humans, there are general tendencies for preferring simpler and more general explanations [84]. However, as described earlier, there are individual differences in which explanations are considered high quality. The context of the decision and the type of decision being made can influence this as well. Humans do not make decisions in isolation of other factors [64]. Without conscious awareness, people incorporate irrelevant information into a variety of decisions such as first impressions, personality trait judgments, and jury decisions [33, 48, 132, 133]. Even when provided identical information, the context, a person's biases, and the way in which information is presented influences decisions [10, 25, 28, 36, 54, 62, 100, 136]. Considering these human factors within the context of explainable AI has only just begun.

To succeed in explainable AI, the community needs to study the interface between humans and AI systems. Human-machine collaborations have shown to be highly effective in terms of accuracy [98]. There may be similar breakthroughs for AI explainability in human-machine collaborations. The principles defined here provide guidance and a philosophy for driving explainable AI toward a safer world by giving users a deeper under-

standing into a system's output. Meaningful and accurate explanations empower users to apply this information to adapt their behavior and/or appeal decisions. For developers and auditors, explanations equip them with the ability to improve, maintain, and deploy systems as appropriate. Explainable AI contributes to the safe operation and trust of multiple facets of complex AI systems. The common framework and definitions under the four principles facilitate the evolution of explainable AI methods necessary for complex, real-world systems.

Acknowledgments

The authors thank Harold Booth, John Libert, Reva Schwartz, Rachael Sexton, Brian Stanton, Craig Watson, and Jesse Zhang for their insightful comments and discussions. We thank everyone who responded to our call and submitted comments to the draft version of this paper. We thank the panelists and participants of the NIST Explainable AI Workshop for their discussions, insight, and thought provoking commentary. These were all essential for shaping and improving the paper and the future directions of this work.

References

[1] General Data Protection Regulation (GDPR), 2018.

[2] Amina Adadi and Mohammed Berrada. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2870052.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 9505–9515. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8160-sanity-checks-for-saliency-maps.pdf.

[4] Gael Aglin, Siegfried Nijssen, and Pierre Schaus. Learning Optimal Decision Trees Using Caching Branch-and-Bound Search. 2020. URL https://dial.uclouvain.be/pr/boreal/object/boreal:223390.

[5] Ulrich Aivodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170, May 2019. URL http://proceedings.mlr.press/v97/aivodji19a.html. ISSN: 1938-7228 Section: Machine Learning.

[6] David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7775–7784. Curran Associates, Inc., 2018. URL http://papers.nips.cc/paper/8003-towards-robust-interpretability-with-self-explaining-neural-networks.pdf.

[7] Andrew Anderson, Jonathan Dodge, Amrita Sadarangani, Zoe Juozapaitis, Evan Newman, Jed Irvine, Souti Chattopadhyay, Matthew Olson, Alan Fern, and Margaret Burnett. Mental models of mere mortals with explanations of reinforcement learning. *ACM Trans. Interact. Intell. Syst.*, 10(2), May 2020. ISSN 2160-6455. doi: 10.1145/3366485. URL https://doi.org/10.1145/3366485.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, May 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[9] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012. URL http://www.sciencedirect.com/science/article/pii/S1566253519308103.

[10] Marianne Bertrand and Sendhil Mullainathan. Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, 2004. doi: 10.4324/9780429499821-53.

[11] Dimitris Bertsimas and Jack Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, July 2017. ISSN 1573-0565. doi: 10.1007/s10994-017-5633-9. URL https://doi.org/10.1007/s10994-017-5633-9.

[12] Umang Bhatt, José M. F. Moura, and Adrian Weller. Evaluating and Aggregating Feature-based Model Explanations. volume 3, pages 3016–3022, July 2020. doi: 10.24963/ijcai.2020/417. URL https://www.ijcai.org/proceedings/2020/417. ISSN: 1045-0823.

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

[14] Markus Bindemann, Janice Attard, and Robert A. Johnston. Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, 1(1), 2014. ISSN 23311908. doi: 10.1080/23311908.2014.986903. URL http://dx.doi.org/10.1080/23311908.2014.986903.

[15] Anna K. Bobak, Viktoria R Mileva, and Peter JB Hancock. Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, page 174702181877614, 2018. ISSN 1747-0218. doi: 10.1177/1747021818776145.

[16] Charles F Bond and Bella M DePaulo. Accuracy of Deception Judgments Characterizations of Deception. *Personality and Social Psychology Review*, 10(3):214–234, 2006.

[17] David A. Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence. NISTIR 8367, National Institute of Standards and Technology, 2021.

[18] David A. Broniatowski and Valerie F. Reyna. A formal model of fuzzy-trace theroy: Variations on framing effects and the Allais paradox. *Decision (Wash D C)*, 5(4): 205–252, 2018. doi: 10.1037/dec0000083.

[19] Rich Caruana, Scott Lundberg, Marco Tulio Ribeiro, Harsha Nori, and Samuel Jenkins. Intelligible and Explainable Machine Learning: Best Practices and Practical Challenges. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3511–3512. Association for Computing Machinery, New York, NY, USA, August 2020. ISBN 978-1-4503-7998-4. URL https://doi.org/10.1145/3394486.3406707.

[20] C. Chen, K. Lin, Cynthia Rudin, Y. Shaposhnik, S. Wang, and T. Wang. An explainable model for credit risk performance. Technical report, 2018. URL https://users.cs.duke.edu/{~}cynthia/docs/DukeFICO2018Documentation.pdf.

[21] M. Chi. Two approaches to the study of experts' characteristics. In K. Ericsson, N. Charness, P. Feltovich, and R. Hoffman, editors, *The Cambridge Handbook of Expertise and Expert Performance*, chapter 2, pages 21–30. Cambridge University Press, Cambridge, 2006. doi: 10.1017/CBO9780511816796.002.

[22] Chief Finalcial Officers Council and Performance Improvement Council. Playbook: Enterprise Risk Management for the U.S. Federal Government. Technical report, July 2016. URL https://www.cfo.gov/wp-content/uploads/2016/07/FINAL-ERM-Playbook.pdf.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[24] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *European Conference on Artificial Intelligence*, 2020.

[25] Jennifer L. Doleac and Luke C.D. Stein. The visible hand: Race and online market outcomes. *The Economic Journal*, 123(572):F469–F492, 2013. doi: 10.1111/ecoj.12082. URL http://www.jstor.com/stable/42919259.

[26] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[27] F. K. Došilović, M. Brčić, and N. Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 0210–0215, May 2018. doi: 10.23919/MIPRO.2018.8400040.

[28] Gary Edmond, Alice Towler, Bethany Growns, Gianni Ribeiro, Bryan Found, David White, Kaye Ballantyne, Rachel A. Searston, Matthew B. Thompson, Jason M. Tangen, Richard I. Kemp, and Kristy Martire. Thinking forensics: Cognitive science for forensic practitioners. *Science and Justice*, 57(2):144–154, 2017. ISSN 18764452.

doi: 10.1016/j.scijus.2016.11.005. URL http://dx.doi.org/10.1016/j.scijus.2016.11.005.

[29] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, January 2015.

[30] Marte Fallshore and Jonathan W. Schooler. Verbal Vulnerability of Perceptual Expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21 (6):1608–1623, 1995. ISSN 02787393. doi: 10.1037/0278-7393.21.6.1608.

[31] Board of Governors Federal Reserve System. SR 11-7: Guidance on Model Risk Management. Supervision and Regulation Letters SR 11-7, The Federal Reserve System, April 2011. URL https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm.

[32] Tom Ferguson. *Game Theory*. Second edition, 2014. URL https://www.math.ucla.edu/~tom/Game_Theory/Contents.html.

[33] Heather D. Flowe and Joyce E. Humphries. An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology*, 25(2):265–273, 2011. ISSN 08884080. doi: 10.1002/acp.1673.

[34] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, pages 80–89, 2018. doi: 10.1109/DSAA.2018.00018.

[35] Google Inc. Facets, 2019. URL https://pair-code.github.io/facets/.

[36] Google LLC. AI Explanations Whitepaper. pages 1–28, 2019.

[37] United States Government Accountability Office. Report to the Committee on Oversight and Government Reform, House of Representatives. Technical Report GAO-17-63, December 2016. URL https://www.gao.gov/products/gao-17-63.

[38] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys (CSUR)*, 51(5):93:1–93:42, August 2018. ISSN 0360-0300. doi: 10.1145/3236009. URL https://doi.org/10.1145/3236009.

[39] Lars Hall, Petter Johansson, Betty Tärning, Sverker Sikström, and Thérèse Deutgen. Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, 117(1):54–61, 2010. ISSN 00100277. doi: 10.1016/j.cognition.2010.06.010. URL http://dx.doi.org/10.1016/j.cognition.2010.06.010.

[40] Lars Hall, Petter Johansson, and Thomas Strandberg. Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE*, 7(9), 2012. ISSN 19326203. doi: 10.1371/journal.pone.0045457.

[41] Patrick Hall, Navdeep Gill, and Nicholas Schmidt. Proposed guidelines for the responsible use of explainable machine learning, 2019.

[42] Nigel Harvey. Confidence in judgment. *Trends in Cognitive Sciences*, 1(2):78–82, 1997. ISSN 13646613. doi: 10.1016/S1364-6613(97)01014-0.

[43] Peter Hase and Mohit Bansal. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL https://www.aclweb.org/anthology/2020.acl-main.491.

[44] Michael Hind. Explaining Explainable AI. *XRDS*, 25(3):16–19, April 2019. ISSN 1528-4972. doi: 10.1145/3313096. URL http://doi.acm.org/10.1145/3313096.

[45] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. Metrics for Explainable AI: Challenges and Prospects. *arXiv:1812.04608 [cs]*, February 2019. URL http://arxiv.org/abs/1812.04608. arXiv: 1812.04608.

[46] Andreas Holzinger, André Carrington, and Heimo Müller. Measuring the Quality of Explanations: The System Causability Scale (SCS). *KI - Künstliche Intelligenz*, 34(2):193–198, June 2020. ISSN 1610-1987. doi: 10.1007/s13218-020-00636-z. URL https://doi.org/10.1007/s13218-020-00636-z.

[47] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A Benchmark for Interpretability Methods in Deep Neural Networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 9737–9748. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

[48] Ying Hu, Connor J. Parde, Matthew Q. Hill, Naureen Mahmood, and Alice J. O'Toole. First Impressions of Personality Traits From Body Shapes. *Psychological Science*, 29(12):1969–1983, 2018. ISSN 14679280. doi: 10.1177/0956797618799300.

[49] IBM Research. Trusting AI, Accessed July 8, 2020. URL https://www.research.ibm.com/artificial-intelligence/trusted-ai/.

[50] Information Commissioner's Office and The Alan Turing Institute. Explaining decisions made with AI, 2020. URL https://ico.org.uk/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-ai/.

[51] G. Jackson, D. H. Kaye, C. Neumann, A. Ranadive, and V. F. Reyna. Communicating the Results of Forensic Science Examinations. Technical report, 2015.

[52] Natalie Japkowicz and Mohak Shah. *Evaluating Learning Algorithms A Classification Perspective*. Cambridge University Press, June 2014. URL http://www.cambridge.org/us/academic/subjects/computer-science/pattern-recognition-and-machine-learning/evaluating-learning-algorithms-classification-perspective.

[53] Petter Johansson, Lars Hall, Sverker Sikström, and Andreas Olsson. Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310 (5745):116–119, 2005. ISSN 00368075. doi: 10.1126/science.1111709.

[54] Saul M. Kassin, Itiel E. Dror, and Jeff Kukucka. The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in*

Memory and Cognition, 2(1):42–52, 2013. ISSN 22113681. doi: 10.1016/j.jarmac.2013.01.001. URL http://dx.doi.org/10.1016/j.jarmac.2013.01.001.

[55] Frank C. Keil. Explanation and understanding. *Annual Review of Psychology*, 57: 227–254, 2006. doi: 10.1146/annurev.psych.57.102904.190100. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3624763/pdf/nihms412728.pdf.

[56] Been Kim, Cynthia Rudin, and Julie A Shah. The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1952–1960. Curran Associates, Inc., 2014. URL http://papers.nips.cc/paper/5313-the-bayesian-case-model-a-generative-approach-for-case-based-reasoning-and-prototype-classi pdf.

[57] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (Un)reliability of Saliency Methods. In Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller, editors, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Lecture Notes in Computer Science, pages 267–280. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_14. URL https://doi.org/10.1007/978-3-030-28954-6_14.

[58] Pang Wei Koh and Percy Liang. Understanding Black-Box Predictions via Influence Functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 1885–1894. JMLR.org, 2017. event-place: Sydney, NSW, Australia.

[59] Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felton, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. Accountable Algorithms. *University of Pennsylvania Law Review*, pages 633–705, 2017.

[60] Justin Kruger and David Dunning. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134, 1999.

[61] D. Richard Kuhn, Raghu Kacker, Yu Lei, and Dimitris E. Simos. Combinatorial Methods for Explainable AI. In *IWCT 2020*, March 2020. URL https://conf.researchr.org/details/iwct-2020/iwct-2020-papers/20/Combinatorial-Methods-for-Explainable-AI. Library Catalog: conf.researchr.org.

[62] Jeff Kukucka, Saul M. Kassin, Patricia A. Zapf, and Itiel E. Dror. Cognitive Bias and Blindness: A Global Survey of Forensic Science Examiners. *Journal of Applied Research in Memory and Cognition*, 6(4):452–459, 2017. ISSN 22113681. doi: 10.1016/j.jarmac.2017.09.001. URL http://dx.doi.org/10.1016/j.jarmac.2017.09.001.

[63] Chan Kulatunga-Moruzi, Lee R. Brooks, and Geoffrey R. Norman. Using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30(3):563–572, 2004. ISSN 02787393. doi: 10.1037/0278-7393.30.3.563.

[64] Kestutis Kveraga, Avniel S. Ghuman, and Moshe Bar. Top-down prediction in the

cognitive brain. *Brain and cognition*, 65(2):145–168, 2007. doi: 10.1016/j.bandc.2007.06.007.

[65] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. An Evaluation of the Human-Interpretability of Explanation. *arXiv:1902.00006 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1902.00006. arXiv: 1902.00006.

[66] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):59–67, October 2019. URL https://aaai.org/ojs/index.php/HCOMP/article/view/5280.

[67] Vivian Lai and Chenhao Tan. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. November 2018. doi: 10.1145/3287560.3287590. URL https://arxiv.org/abs/1811.07901v4.

[68] Hima Lakkaraju, Julius Adebayo, and Sameer Singh. Explaining Maching Learning Predictions: State-of-the-art, Challenges, and Opportunities, December 2020. URL https://explainml-tutorial.github.io/neurips20.

[69] Himabindu Lakkaraju and Osbert Bastani. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 79–85, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375833. URL https://doi.org/10.1145/3375627.3375833.

[70] Himabindu Lakkaraju and Cynthia Rudin. Learning Cost-Effective and Interpretable Treatment Regimes. In *Artificial Intelligence and Statistics*, pages 166–175, April 2017. URL http://proceedings.mlr.press/v54/lakkaraju17a.html.

[71] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 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, pages 1675–1684, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939874. URL https://doi.org/10.1145/2939672.2939874. event-place: San Francisco, California, USA.

[72] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 131–138, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6324-2. doi: 10.1145/3306618.3314229. URL https://doi.org/10.1145/3306618.3314229. event-place: Honolulu, HI, USA.

[73] Jared LeClerc and Susan Joslyn. The cry wolf effect and weather-related decision making. *Risk analysis*, 35(3):385–395, 2015.

[74] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke

prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, September 2015. ISSN 1932-6157, 1941-7330. doi: 10.1214/15-AOAS848. URL https://projecteuclid.org/euclid.aoas/1446488742.

[75] Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*, April 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17082.

[76] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. *Conference on Human Factors in Computing Systems - Proceedings*, pages 2119–2128, 2009. doi: 10.1145/1518701.1519023.

[77] Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and Scalable Optimal Sparse Decision Trees. In *Proceedings of the International Conference on Machine Learning*, volume 1, 2020. URL https://proceedings.icml.cc/paper/2020/hash/8a1ee9f2b7abe6e88d1a479ab6a42c5e-Abstract.html.

[78] Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy*, 23(1):18, January 2021. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[79] Zachary C Lipton. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43, 2018.

[80] Tania Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470, October 2006. ISSN 1364-6613. doi: 10.1016/j.tics.2006.08.004. URL http://www.sciencedirect.com/science/article/pii/S1364661306002117.

[81] José Marcio Luna, Efstathios D. Gennatas, Lyle H. Ungar, Eric Eaton, Eric S. Diffenderfer, Shane T. Jensen, Charles B. Simone, Jerome H. Friedman, Timothy D. Solberg, and Gilmer Valdes. Building more accurate decision trees with the additive tree. *Proceedings of the National Academy of Sciences*, 116(40):19887–19893, October 2019.

[82] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[83] Oisin Mac Aodha, Shihan Su, Yuxin Chen, Pietro Perona, and Yisong Yue. Teaching Categories to Human Learners with Visual Explanations. February 2018. URL https://arxiv.org/abs/1802.06924v1.

[84] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007. URL http://www.sciencedirect.com/science/article/pii/S0004370218305988.

[85] Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. Model reconstruction from model explanations, 2018. URL https://arxiv.org/abs/1807.05185v1.

[86] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 220–229, New York, NY, USA, January 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

[87] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv:1811.11839 [cs]*, August 2020. URL http://arxiv.org/abs/1811.11839. arXiv: 1811.11839.

[88] Christoph Molnar. Interpretable Machine Learning, 2018.

[89] Christoph Molnar. *Interpretable Machine Learning*. @ChristophMolnar, online edition edition, April 2019. URL https://christophm.github.io/interpretable-ml-book/.

[90] Shane T. Mueller, Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *arXiv:1902.01876 [cs]*, February 2019. URL http://arxiv.org/abs/1902.01876. arXiv: 1902.01876.

[91] Shane T. Mueller, Elizabeth S. Veinott, Robert R. Hoffman, Gary Klein, Lamia Alam, Tauseef Mamun, and William J. Clancey. Principles of explanation in human-ai systems. *arXiv preprint arXiv:2102.04972*, 2021. URL https://arxiv.org/abs/2102.04972.

[92] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv:1802.00682 [cs]*, February 2018. URL http://arxiv.org/abs/1802.00682. arXiv: 1802.00682.

[93] Richard E Nisbett, Timothy Decamp Wilson, Michael Kruger, Lee Ross, Amos Indeed, Nancy Bellows, Dorwin Cartwright, Alvin Goldman, Sharon Gurwitz, Ronald Lemley, Harvey London, and Hazel Markus. Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 1977.

[94] NIST Standards Inclusivity Effort Team. Guidance for NIST staff on using inclusive language in documentary standards, NIST Interagency or Internal report 8366, 2021.

[95] Stuart Oskamp. Overconfidence in case-study judgments. *Journal of Consulting Psychology*, 29(3):261–265, 1965. ISSN 00958891. doi: 10.1037/h0022125.

[96] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. PAMI*, 22:1090–1104, October 2000.

[97] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. PAMI*, 32(5):831–846, 2010.

[98] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey

Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, 2018.

[99] P.J. Phillips, K. W. Bowyer, P. J. Flynn, X. Liu, and W. T. Scruggs. The Iris Challenge Evaluation 2005. In *Second IEEE International Conference on Biometrics: Theory, Applications, and Systems*, 2008.

[100] Rüdiger F. Pohl, editor. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. Psychology Press, 2004.

[101] Forough Poursabzi-Sangdeh, Daniel G. Goldstein, Jake M. Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*, November 2019. URL http://arxiv.org/abs/1802.07810. arXiv: 1802.07810.

[102] Emily Pronin. The introspection illusion. In *Advances in experimental social psychology*, pages 1–67. Elsevier, 2009.

[103] Mark A Przybocki, Alvin F Martin, and Audrey N Le. Nist speaker recognition evaluations utilizing the mixer corpora—2004, 2005, 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):1951–1959, 2007.

[104] Zhongang Qi, Saeed Khorram, and Li Fuxin. Visualizing Deep Networks by Optimizing with Integrated Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11890–11898, April 2020. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v34i07.6863. URL http://arxiv.org/abs/1905.00954. arXiv: 1905.00954.

[105] Dheeraj Rajagopal, Vidhisha Balachandran, Eduard Hovy, and Yulia Tsvetkov. SelfExplain: A Self-Explaining Architecture for Neural Text Classifiers. *arXiv:2103.12279 [cs]*, March 2021. URL http://arxiv.org/abs/2103.12279. arXiv: 2103.12279.

[106] Kaivalya Rawal and Himabindu Lakkaraju. Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses. In *Advances in Neural Information Processing Systems*, volume 33, pages 12187–12198. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/8ee7730e97c67473a424ccfeff49ab20-Abstract.html.

[107] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust you?" Explaining the Predictions of Any Classifier. In *KDD 2016: Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, August 2016. ACM. URL https://www.kdd.org/kdd2016/subtopic/view/why-should-i-trust-you-explaining-the-predictions-of-any-classifier.

[108] Allyson Rice, P. Jonathon Phillips, and Alice J. O'Toole. The role of the face and body in unfamiliar person identification. *Applied Cognitive Psychology*, 27:761–768, 2013.

[109] John Roach. Microsoft responsible machine learning capabilities build trust in AI

systems, developers say, Accessed July 29, 2020. URL https://blogs.microsoft.com/ai/azure-responsible-machine-learning/.

[110] Marko Robnik-Šikonja and Marko Bohanec. Perturbation-Based Explanations of Prediction Models. In Jianlong Zhou and Fang Chen, editors, *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*, Human–Computer Interaction Series, pages 159–175. Springer International Publishing, Cham, 2018. ISBN 978-3-319-90403-0. doi: 10.1007/978-3-319-90403-0_9. URL https://doi.org/10.1007/978-3-319-90403-0_9.

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

[112] Cynthia Rudin and Joanna Radin. Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. *Harvard Data Science Review*, 1(2), 2019.

[113] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021. URL https://arxiv.org/abs/2103.11251.

[114] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero. The 2016 nist speaker recognition evaluation. In *Interspeech*, pages 1353–1357, 2017.

[115] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28 (11):2660–2673, November 2017. ISSN 2162-2388. doi: 10.1109/TNNLS.2016.2599820. Conference Name: IEEE Transactions on Neural Networks and Learning Systems.

[116] Philipp Schmidt and Felix Biessmann. Quantifying Interpretability and Trust in Machine Learning Systems. *arXiv:1901.08558 [cs, stat]*, January 2019. URL http://arxiv.org/abs/1901.08558. arXiv: 1901.08558.

[117] Jonathan W. Schooler and Tonya Y. Engstler-Schooler. Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1): 36–71, 1990. ISSN 00100285. doi: 10.1016/0010-0285(90)90003-M.

[118] Jonathan W. Schooler, Stellan Ohlsson, and Kevin Brooks. Thoughts Beyond Words: When Language Overshadows Insight. *Journal of Experimental Psychology: General*, 122(2):166–183, 1993. ISSN 00963445. doi: 10.1037/0096-3445.122.2.166.

[119] Reva Schwartz, Leann Down, Adam Jonas, and Elham Tabassi. A proposal for identifying and managing bias in artificial intelligence. Draft NIST Special Publication 1270, National Institute of Standards and Technology, 2021.

[120] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International*

Conference on Computer Vision, pages 618–626, 2017.

[121] Keng Siau and Weiyu Wang. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53, 2018. ISSN 24753742.

[122] Leon Sixt, Maximilian Granz, and Tim Landgraf. When Explanations Lie: Why Many Modified BP Attributions Fail. December 2019. URL https://arxiv.org/abs/1912.09818v6.

[123] Dylan Slack, Sorelle A. Friedler, Carlos Scheidegger, and Chitradeep Dutta Roy. Assessing the Local Interpretability of Machine Learning Models. *arXiv:1902.03501 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1902.03501. arXiv: 1902.03501.

[124] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

[125] Chun Siong Soon, Marcel Brass, Hans Jochen Heinze, and John Dylan Haynes. Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545, 2008. ISSN 10976256. doi: 10.1038/nn.2112.

[126] Siegfried Ludwig Sporer, Steven Penrod, Don Read, and Brian Cutler. Choosing, Confidence, and Accuracy: A Meta-Analysis of the Confidence-Accuracy Relation in Eyewitness Identification Studies. *Psychological Bulletin*, 118(3):315–327, 1995. ISSN 00332909. doi: 10.1037/0033-2909.118.3.315.

[127] Brian Stanton and Theodore Jensen. Trust and artificial intelligence. Draft NISTIR 8332, National Institute of Standards and Technology, 2021.

[128] Kevin Stine, Stephen Quinn, Gregory Witte, and Robert Gardner. Integrating Cybersecurity and Enterprise Risk Management (ERM). Technical Report NIST Internal or Interagency Report (NISTIR) 8286, National Institute of Standards and Technology, October 2020. URL https://csrc.nist.gov/publications/detail/nistir/8286/final.

[129] Leanne ten Brinke, Dayna Stimson, and Dana R. Carney. Some Evidence for Unconscious Lie Detection. *Psychological Science*, 2014.

[130] Leanne ten Brinke, Kathleen D. Vohs, and Dana R. Carney. Can Ordinary People Detect Deception After All? *Trends in Cognitive Sciences*, 20(8):579–588, 2016. ISSN 13646613. doi: 10.1016/j.tics.2016.05.012. URL http://linkinghub.elsevier.com/retrieve/pii/S1364661316300547.

[131] The Royal Society. Explainable AI: the basics policy briefing, 2019. URL https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

[132] Alexander Todorov. *Face value: The irresistible influence of first impressions*. Princeton University Press, 2017.

[133] Alexander Todorov, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. Inferences of competence from faces predict election outcomes. *Science (New York,*

N.Y.), 308(5728):1623–6, 6 2005. ISSN 1095-9203. doi: 10.1126/science.1110589. URL http://www.ncbi.nlm.nih.gov/pubmed/15947187.

[134] Ehsan Toreini, Mhairi Aitken, Kovila Coopamootoo, Karen Elliot, Carlos Gonzalez-Zelaya, and Aad van Moorsel. The relationship between trust in AI and trustworthy machine learning technologies. In *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, Barcelona, Spain, 2020. doi: 10.1145/3351095.3372834.

[135] Alice Towler, David White, and Richard I Kemp. Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, 23(1):47, 2017.

[136] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981. doi: 10.1126/science.7455683.

[137] Amos Tversky and Eldar Shafir. The Disjunction Effect in Choice Under Uncertainty. *Psychological Science*, 3(5):305–309, 1992. ISSN 14679280. doi: 10.1111/j.1467-9280.1992.tb00678.x.

[138] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 978-1-4503-6125-5. doi: 10.1145/3287560.3287566. URL https://doi.org/10.1145/3287560.3287566. event-place: Atlanta, GA, USA.

[139] Ellen M Voorhees. System Explanations : A Cautionary Tale. In *ACM CHI Workshop on Operationalizing Human-Centered Perspectives in Explainable AI*, 2021.

[140] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[141] Adrian Weller. Transparency: Motivations and challenges. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 23–40. Springer, 2019.

[142] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. doi: 10.1109/TVCG.2019.2934619.

[143] Timothy D. Wilson and Yoav Bar-Anan. The unseen mind. *Science*, 321(5892): 1046–1047, 2008. ISSN 00368075. doi: 10.1126/science.1163029.

[144] Timothy D. Wilson and Johnathan Schooler. Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2):181–192, 1991. ISSN 0003-066X. doi: 10.1037/h0021466.

[145] John T. Wixted, Laura Mickes, and Ronald P. Fisher. Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science*, 13(3):324–335, 2018. ISSN 17456924. doi: 10.1177/1745691617734878.

[146] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeff Warshaw. A qualitative exploration of perceptions of algorithmic fairness. *Conference on Human Factors in Computing Systems - Proceedings*, 2018-April:1–14, 2018. doi: 10.1145/3173574.3174230.

[147] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I. Inouye, and Pradeep K. Ravikumar. On the (In)fidelity and Sensitivity of Explanations. pages 10967–10978, 2019. URL http://papers.neurips.cc/paper/9278-on-the-infidelity-and-sensitivity-of-explanations.

[148] Nick Yeung and Christopher Summerfield. Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1310–1321, 2012. ISSN 14712970. doi: 10.1098/rstb.2011.0416.

[149] Qingyuan Zhao and Trevor Hastie. Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 0(0):1–10, June 2019. ISSN 0735-0015. doi: 10.1080/07350015.2019.1624293. URL https://doi.org/10.1080/07350015.2019.1624293.

[150] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[151] Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics*, 10(5):593, January 2021. doi: 10.3390/electronics10050593. URL https://www.mdpi.com/2079-9292/10/5/593. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[152] Xingchen Zhou and Rob Jenkins. Dunning–Kruger effects in face perception. *Cognition*, 203(January), 2020. ISSN 18737838. doi: 10.1016/j.cognition.2020.104345.

[153] Luisa M. Zintgraf, Taco S. Cohen, Tameem Adel, and Max Welling. Visualizing Deep Neural Network Decisions: Prediction Difference Analysis. February 2017. URL https://arxiv.org/abs/1702.04595v1.