Draft NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
David A. Broniatowski
Mark A. Przybocki

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

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

Draft NISTIR 8312

Four Principles of Explainable Artificial Intelligence

P. Jonathon Phillips
Carina A. Hahn
Peter C. Fontana
Information Access Division
Information Technology Laboratory

David A. Broniatowski
Information Technology Laboratory

Mark A. Przybocki
Information Access Division
Information Technology Laboratory

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

August 2020

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Interagency or Internal Report 8312
24 pages (August 2020)

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

Public comment period: August 17, 2020 through October 15, 2020

National Institute of Standards and Technology
100 Bureau Drive (Mail Stop 8940) Gaithersburg, Maryland 20899-2000
Email: explainable-AI@nist.gov

All comments will be made public and are subject to release under the Freedom of Information Act (FOIA).

Additional information on submitting comments can be found at https://www.nist.gov/topics/artificial-intelligence/ai-foundational-research-explainability.

Call for Patent Claims

This public review includes a call for information on essential patent claims (claims whose use would be required for compliance with the guidance or requirements in this Information Technology Laboratory (ITL) draft publication). Such guidance and/or requirements may be directly stated in this ITL Publication or by reference to another publication. This call also includes disclosure, where known, of the existence of pending U.S. or foreign patent applications relating to this ITL draft publication and of any relevant unexpired U.S. or foreign patents.

ITL may require from the patent holder, or a party authorized to make assurances on its behalf, in written or electronic form, either:

a) assurance in the form of a general disclaimer to the effect that such party does not hold and does not currently intend holding any essential patent claim(s); or

b) assurance that a license to such essential patent claim(s) will be made available to applicants desiring to utilize the license for the purpose of complying with the guidance or requirements in this ITL draft publication either:

  i. under reasonable terms and conditions that are demonstrably free of any unfair discrimination; or

  ii. without compensation and under reasonable terms and conditions that are demonstrably free of any unfair discrimination.

Such assurance shall indicate that the patent holder (or third party authorized to make assurances on its behalf) will include in any documents transferring ownership of patents subject to the assurance, provisions sufficient to ensure that the commitments in the assurance are binding on the transferee, and that the transferee will similarly include appropriate provisions in the event of future transfers with the goal of binding each successor-in-interest.

The assurance shall also indicate that it is intended to be binding on successors-in-interest regardless of whether such provisions are included in the relevant transfer documents.

Such statements should be addressed to: explainable-AI@nist.gov

Abstract

We introduce four principles for explainable artificial intelligence (AI) that comprise the fundamental properties for explainable AI systems. They 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; trustworthy AI.

Table of Contents

List of Figures

1. Introduction

With recent advances in artificial intelligence (AI), AI systems have become components of high-stakes decision processes. 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, provide information about the logic behind those decisions[1] and the desire to create trustworthy AI [30, 76, 89].

Based on these calls for explainable systems [40], 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 oneself and to society [102]. This may slow societal acceptance and adoption of the technology, as members of the general public oftentimes place the burden of meeting societal goals on manufacturers and programmers themselves [27, 102]. Therefore, in terms of societal acceptance and trust, developers of AI systems may need to consider that multiple attributes of an AI system can influence public perception of the system.

Explainable AI is one of several properties that characterize trust in AI systems [83, 92]. 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. For this paper, we state four principles encompassing the core concepts of explainable AI. These are informed by research from the fields of computer science, engineering, and psychology. In considering aspects across these fields, this report provides a set of contributions. First, we articulate the four principles of explainable AI. From a computer science perspective, we place existing explainable AI algorithms and systems into the context of these four principles. From a psychological perspective, we investigate how well people's explanations follow our four principles. This provides a baseline comparison for progress in 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. Rather, we outline a set of principles that organize and review existing work in explainable AI and guide future research directions for the field. These principles support the foundation of policy considerations, safety, acceptance by society, and other aspects of AI technology.

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

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

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.

Before proceeding with the principles, we need to define a key term, the *output* of an AI system. The output is the result of a query to an AI system. The output of a system varies by task. A loan application is an example where 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.

Briefly, our four principles of explainable AI are:

Explanation: Systems deliver accompanying evidence or reason(s) for all outputs.

Meaningful: Systems provide explanations that are understandable to individual users.

Explanation Accuracy: The explanation correctly reflects the system's process for generating the output.

Knowledge Limits: The system only operates under conditions for which it was designed or when the system reaches a sufficient confidence in its output.

These are defined and contextualized in more detail below.

2.1 Explanation

The *Explanation* principle obligates AI systems to supply evidence, support, or reasoning for each output. By itself, this principle does not require that the evidence be correct, informative, or intelligible; it merely states that a system is capable of providing an explanation. A body of ongoing work currently seeks to develop and validate explainable AI methods. An overview of these efforts is provided in Section 5. A variety of strategies and tools are currently being deployed and developed. This principle does not impose any metric of quality on those explanations. The Meaningful and Explanation Accuracy principles provide a framework for evaluating explanations.

2.2 Meaningful

A system fulfills the *Meaningful* principle if the recipient understands the system's explanations. Generally, this principle is fulfilled if a user can understand the explanation, and/or it is useful to complete a task. This principle does not imply that the explanation is one size fits all. Multiple groups of users for a system may require different explanations. The Meaningful principle allows for explanations which are tailored to each of the user groups. Groups may be defined broadly as the developers of a system vs. end-users of a system; lawyers/judges vs. juries; etc. The goals and desiderata for these groups may vary. For example, what is meaningful to a forensic practitioner may be different than what is meaningful to a juror [31].

This principle also allows for tailored explanations at the level of the individual. Two humans viewing the same AI system's output will not necessarily interpret it the same way for a variety of reasons. One reason is that a person's prior knowledge and experiences influence their decisions [45]. Another reason is that psychological differences among people may influence how they interpret an explanation and what type of explanations they find meaningful [10, 61]. Thus, different users may take different meanings from identical AI explanations. The tailoring of an explanation to user groups and individuals may not be static over time. As people gain experience with a task, what they consider a meaningful explanation will likely change [10, 35, 57, 72, 73]. Therefore, meaningfulness is influenced by a combination of the AI system's explanation and a person's prior knowledge, experiences, and mental processes.

All of the factors that influence meaningfulness contribute to the difficulty in modeling the interface between AI and humans. Developing systems that produce meaningful explanations need to account for both computational and human factors [22, 58].

2.3 Explanation Accuracy

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

Explanation accuracy is a distinct concept from decision accuracy. For decision tasks, 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. Researchers in AI have developed standard measures of algorithm and system accuracy [13, 18, 33, 64–66, 71, 79]. While there exist these established decision accuracy metrics, researchers are in the process of developing performance metrics for explanation accuracy [2, 16, 97].

Similarly to the Meaningful principle, this principle allows for different explanation accuracy metrics for different groups and individuals. Some users will require simple explanations that succinctly focus on the critical point(s) but lack nuances that are necessary to completely characterize the algorithm's process for generating its output. However, these nuances may only be meaningful to experts. This highlights the point that explanation accuracy and meaningfulness need not overlap. A detailed explanation may be highly accurate but sacrifice how meaningful it is to certain audiences. Overall, a system may be considered more explainable if it can generate more than one type of of explanation. Because of these different levels of explanation, the metrics used to evaluate the accuracy of an explanation may not be universal or absolute.

2.4 Knowledge Limits

The previous principles implicitly assume that a system is operating within its knowledge limits. This *Knowledge Limits* principle states that systems identify cases they were not designed or approved to operate, or 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. The Knowledge Limits Principle can increase trust in a system by preventing misleading, dangerous, or unjust decisions or outputs.

There are two ways a system can reach its knowledge limits. First, the question can be outside the domain of the system. 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 the second way a knowledge limit can be reached, the confidence of the most likely answer may be too low, depending on an internal confidence threshold. For example, for a 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."

3. Types of Explanations

Explanations will vary depending on their consumer. Some explanations will be simple, while others will be detailed and could require training or expertise to fully understand. To illustrate the range of explanation, we describe five categories of explanations that build on the work in the literature [6, 26, 98]. The categories described below were not designed to be exhaustive.

User benefit: This type of explanation is designed to inform a user about an output. For example, the explanation could provide the reason a loan application was approved or denied to the applicant.

Societal acceptance: This type of explanation is designed to generate trust and acceptance by society. For example, if an unexpected output is provided by the system, the explanation may help users understand why this output was generated. It may also provide an increased sense of comfort in the system if the rationale can be provided (e.g., [1]).

Regulatory and compliance: This type of explanation assists with audits for compliance with regulations, safety standards, etc. The audience of the explanation may include a user who requires significant detail (e.g., a safety regulator) and the user interacting with the system (e.g., a developer). Examples may include the developer or auditor of a self-driving car. This may also include explanations to evaluate the output of a forensic examination after an airplane crash.

System development: This type of explanation assists or facilitates developing, improving, debugging, and maintaining of an AI algorithm or system. Consumers of this category includes technical staff, product managers, and executives. This category includes the users requiring significant detail and users interacting with the system. For example, this may include the technical staff debugging a vision algorithm with a Gradient-Weighted Class Activation Mapping (GRAD-CAM) based tool [82].

Owner benefit: This type of explanation benefits the operator of a system. An example is a recommendation system that lists movies or videos to watch and explains the selection based on previous viewed items. A system recommends a movie and explains this choice by stating "here is a movie to watch because you liked these other movies." If the user trusts the explanation, the owner benefits because that person continues watching movies on their service.

Categories of this nature are also discussed in more detail in Bhatt et al. [6], Hall et al. [26], Weller [98]. Bhatt et al. [6] mentions in their use cases that the explanations are usually used by the algorithm developers to debug the models. Bhatt et al. [6] interviews 30 individuals on how their organizations use explainable AI. They use explainable AI in a variety of applications, including object detection and sentiment analysis. Hall et al. [26] proposes best practices on how to use explainable AI algorithms. They summarize their recommendations into implementation guidelines: design explanations to enable understanding, learn how explainable AI can be exploited for nefarious purposes, augment surrogate models with direct explanations, and for high-stakes decisions, provided explanations must be highly interpretable. In Caruana et al. [11], the authors developed an explainable AI model and used it to both determine and explain pneumonia risk in a patient data set and 30-day readmission risk in another patient data set.

From a practical perspective, explanations can be characterized by the amount of time the consumer of the explanation has to respond to the information and the level of detail in an explanation. Figure 1 captures the relationship between time requirements and explanation detail. The horizontal axis represents the *time requirement* a user has to respond to a situation. The time requirement axis addresses situations ranging from those that require immediate responses to those that permit a longer evaluation. The vertical axis represents the *level of detail* in the explanation. This axis addresses situations related to the level of detail the consumer or user will require. At one end of the explanation, an explanation is not required or a simple explanation will be sufficient. For example, in response to an emergency weather alert, the consumer must act immediately, and the explanation needs to be simple and straightforward. A current weather alert from the National Weather Service, "Tornado Warning: Take Action!"[2], operates as both an alert and a simple explanation. The alert is to "Take Action" with the simple explanation of "Tornado Warning." Explanations for debugging could fall at the other end of the time requirement and level of detail spectrum. The explanation could include information on the internal steps of a system, and it

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

could take the audience time to examine the explanation and decide on their next actions. Two additional examples were placed on Figure 1: loan applications and audit of a system. The response to a loan application is generally quick and the explanation provides greater detail than a weather alert. The response time and explanation detail for an audit of a system could be similar to debugging a system.

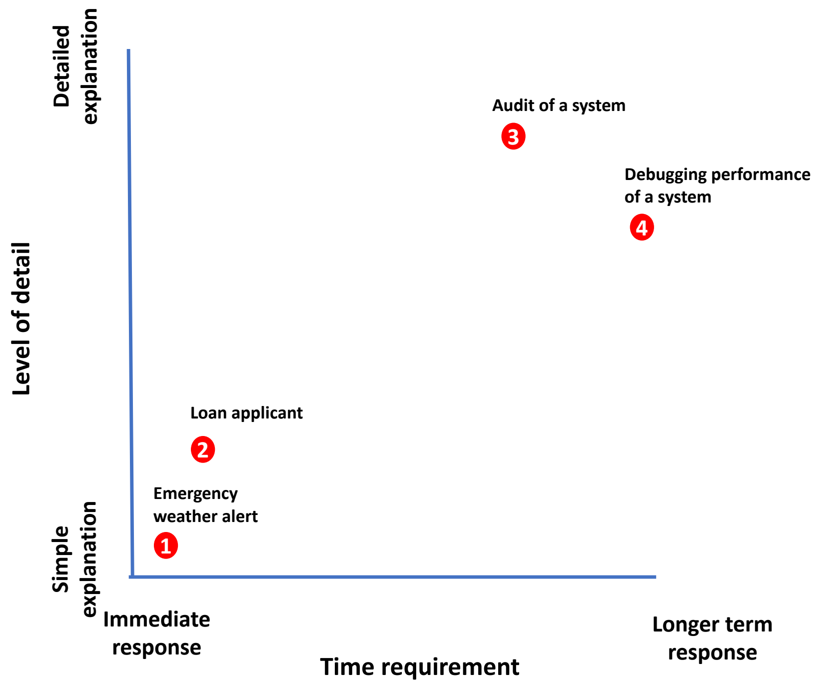

Fig. 1. This figure shows length of response time versus explanation detail. We populate the figure with four illustrative cases: emergency weather alert, loan application, audit of a system, and debugging a system.

Explanations will need to fulfill a variety of requirements and needs, which will depend on the tasks and users. The five categories of explanations illustrate the range and types of explanations and points to the need for flexibility in addressing the scope of systems that require explanations.

4. 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 [16, 22, 53, 77, 78, 98].

Lipton [53] divides explainable techniques into two broad categories: transparent and post-hoc interpretability. Lipton [53] 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 [77] and Rudin and Radin [78] argue that models for high-stakes decision must provide explanations that reveal their inner workings. They claim that deep neural networks are inherently black-boxes and should be avoided for high-stakes decisions.

Wachter et al. [97] 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 [59];" 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. [22] defines a set of concepts for explainable AI and provides an outline of current approaches. In their survey, Gilpin et al. [22] take a similar stance to Rudin [77] and Rudin and Radin [78] in their set of "foundational concepts" for explainability. Similar to the meaningful and explanation accuracy principles in our current work, Gilpin et al. [22] propose that explanations should allow for a trade-off between their interpretability and completeness. However, they state that trade-offs must not obscure key limitations of a system.

Doshi-Velez and Kim [16] 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. This highlights the gap between these principles as a concept and creating metrics and evaluation methods.

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.

5. Overview of Explainable AI Algorithms

Researchers have developed different algorithms to explain AI systems. Sometimes, the algorithms themselves provide the explanation (satisfying Principle 1). The most common of these explanations are *self-explainable models*, where the models themselves are the provided explanation. These models are self-explaining algorithms, where viewing and querying the models provide an explanation. We describe these algorithms in Section 5.1. There are algorithms that provide explanations for themselves without directly providing the model details. One such example is Class Activation Mappings (CAM) [105], which are system-specific explanations that can explain some convolutional neural networks. However, researchers generalized these algorithms so that they can not only explain the original system but also explain other systems. These generalized algorithms form the next two types of explanations: global explainable AI algorithms and per-decision explainable AI algorithms. For instance, GRAD-CAM is a generalization of CAM that can provide the explanation of CAM but to any convolutional neural network [82].

A *global explanation* produces a model that approximates the non-interpretable model. We describe these algorithms in Section 5.2. *Per-decision explanations* provide a separate explanation for each decision. Per-decision explanations are considered *local explanations*. We describe per-decision explanations in Section 5.3. A particular type of per-decision explanation is a *counterfactual*[97], which is an explanation saying "if the input were this new input instead, the system would have made a different decision." In these explanations, although there are often many widely-differing instances that all are counterfactuals, a counterfactual explanation usually provides a single instance. This means that even if there are many different possible ways that the instance could be changed to result in the system providing the decision, only one of those instances is provided as the explanation. The hope is the instance is as similar as possible to the input with the exception that the system makes a different decision. Because counterfactual explanations are per-decision explanations, they are also described in Section 5.3.

Self-explainable models of machine learning systems themselves can be used as global explanations (since the models explain themselves). Likewise, many global explanations (including self-explainable models) can also be used to generate per-decision explanations. The coefficient weights of the features of an input in a regression model and the flow of a decision through a decision tree both serve as per-decision explanations. Models that do not provide an explanation or provide an explanation that a user does not consider meaningful enough will sometimes seek an explanation from an alternate algorithm, thus encouraging the development of global and per-decision explanations. Furthermore, global explanations are harder to generate than per-decision explanations because per-decision explanations only require an understanding of a single decision.

With these explainable algorithms, developers wish for the explanations to be meaningful to users (Principle 2). In the computer science literature this is often labelled as *interpretable*. Often, developers self-proclaim their algorithm explanations to be meaningful. However, others will use measurements such as human simulatability [46], which measure whether a human can correctly take an input and with the model, correctly identify the model's prediction.

Although the explanation accuracy is important (Principle 3), it is often only measured for self-explainable models. For these types of models, the model's decision accuracy (see Section 2.3) is the measure of the explanation accuracy. However, there is limited research measuring explanation accuracy. Adebayo et al. [2] evaluate explanation accuracy of saliency pixel explanations for deep neural networks by measuring the amount the explanation changes relative to how the trained models differ.

To our knowledge there is limited work on developing algorithms that understand their knowledge limits (Principle 4) and declare when a validly-formatted data input is out of the system's scope. However, algorithms often give real-valued outputs rather than hard decisions, which reflect the algorithms' confidences in their predictions.

5.1 Self-Explainable Models

Machine Learning Algorithms include Decision Trees and Linear and Logistic Regression. Although these simple models are explanations themselves, they are often not always accurate, especially if used without much pre-processing. Consequently, there has been work in developing more accurate models that themselves are explanations. Authors developing models will often label these models as interpretable, which we refer to as meaningful. Rudin [77] argues that using meaningful models that explain themselves are the best way to produce explainable models, arguing that separately-produced explanations of black-box models (or even single decisions of black-box models) may not be faithful to what the original model computes. This claim is that explanations often have low explanation accuracy if those explanations are not the models themselves. Although many sources discuss an accuracy-interpretability trade-off, Rudin and Radin [78] disagrees, with the belief that no such trade-off exists for high-stakes decisions.

One line of research works on producing improvements on the standard decision trees, sometimes represented as a nested sequence of "if-then-else" rules, called *decision lists* [47]. In addition to being inaccurate, Lakkaraju et al. [47] claims that the nesting makes the rules hard to interpret, and develops *Decision Sets*, which are a sequence of "if-then" rules with one default "else" at the end, where each clause is a conjunction of conditions. However, Lakkaraju and Rudin [50] produces decision lists with improved accuracy. Lakkaraju et al. [47] measure the interpretability of the decision sets by measuring metrics on the model: the number of rules, the number of the largest rules, the overlap of the rules (how many instances are classified in more than one if-then rule). The last "else" guarantees that every instance is classified. [49] explores decision lists with at most one customized nesting to further improve accuracy while still being meaningful according to their measures. Bertsimas and Dunn [5] produce a variant of decision trees, called *optimal classification trees*, that split on mixed integer constraints involving multiple variables. These trees focus on preserving the meaningfulness of decision trees but greatly improving their classification accuracy. [55] produce another variant of a more accurate decision tree, called an additive tree, that combines elements of decision tress and gradient boosting to produce more accurate trees. A Bayesian variant of decision lists that was studied for meaningfulness is Bayesian Rule Lists [51], where they add a Bayesian credible interval estimate to each decision rule. Bayesian credible intervals are the Bayesian analog to confidence intervals. Kuhn et al. [42] produces a model that tries to find combinations of features that either exclude a class or specifically identify a particular class. Each set of combinations could be viewed as a clause of a decision set rule.

Models, including linear models such as linear and logistic regression are considered to be explanations of system decisions. One interpretation is using the weights of the coefficients to indicate the importance of features. They are sometimes considered inaccurate when the data is not believed to be linear. One measure of the ease of understanding of a regression model is the number of non-zero coefficients. One way to encourage a regression model to limit the number of features is to *regularize* it with the *lasso*, which penalizes the model for using more features [32], incorporating a trade-off for accuracy and meaningfulness in the training objective function. Although this and other regularization strategies are also used to prevent overfitting in many models including deep neural networks, regularization is one technique to make models sparser, and thus believed to be more understandable. Poursabzi-Sangdeh et al. [69] considers regression models meaningful and aims to measure the value the model coefficients provide to human users trying to use the model. Caruana et al. [11] also treats the more general class of these models, Generalized Additive Models with Pairwise Interactions (GA2M), as understandable models and applies them to a healthcare case study.

Another self-explainable algorithm involves learning *prototypes*, or representative samples of each class, to better understand the algorithm. Models learn and produce prototypes. With these prototypes, the model outputs the class as a weighted combination of the prototypes. Although these prototypes do work on tabular data, Kim et al. [38], Li et al. [52] use this approach for classification on image data sets.

5.2 Global Explainable AI Algorithms

Global Explainable AI Algorithms are an approach that treat the AI algorithm as a blackbox that can be queried and produce a model that explains the algorithm. Depending on what the global model is, it can then be used to produce per-decision explanations.

One such global explainable AI Algorithms is SHAP (SHapley Additive exPlanations) [56]. SHAP provides a global per-feature importance for a regression problem by converting it to a coalitional game from game theory. In coalitional games, there are n players that can team up in different ways to form coalitions and share a payoff depending on which players team up (often the total payoff is largest when all players team up). After players receive a payoff, they must divide the payoff between themselves. One way to divide payoffs with desirable mathematical properties is to give each player their Shapley value as their individual payoff. SHAP treats the regression outputs of a system as a coalitional game where the target is the payoff and each feature is a player that either participates in or does not participate in the coalition with the other features for each row. SHAP then computes the Shapley values for each feature, and uses those values as the feature importance values. See [20] for more information on Shapley values and coalitional games.

In deep neural networks, one such global algorithm is TCAV (Testing with Concept Activation Vectors) [107]. 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) [60, 104]. 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 [60]. 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.

5.3 Per-Decision Explainable AI Algorithms

Per-decision explainable AI algorithms take both a black-box model that can be queried and a single decision of that model, and explain why the model made that particular decision. These explanations differ from global explanations in that the explanation is not required to generalize to other decisions.

One such algorithm is LIME (Local Interpretable Model-Agnostic Explainer) [74]. 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 popular type of local explanations are counterfactuals. A *counterfactual* explanation is an alternate system input where the system's decision on that input differs from the provided input. Good counterfactuals answer the question "what is the minimum amount an input would need to change for the system to change its decision on that input?" Wachter et al. [97] measures how good counterfactual explanations are by measuring how far away the counterfactual is from the original data point, measuring this distance as the Manhattan distance of features after normalizing each feature by its median absolute deviation. Ustun et al. [96] develop a counterfactual explanation of logistic (or linear) regression models. Counterfactuals are represented as the amounts of specific features to change. They further refine their counterfactual explanations by distinguishing which features can be changed, which ones cannot, and which ones can only be changed under certain conditions.

An additional local explanation in Koh and Liang [39] takes a decision and produces an estimate of the influence of each training data point on that particular decision.

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) [105]. A popular saliency pixel algorithm that enhanced CAM is GRAD-CAM [82]. GRAD-CAM generalized CAM so that it can explain any convolutional network.

A variety of saliency pixel explanation algorithms are compared on for their explanation accuracy in Adebayo et al. [2].

5.4 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 [48] observes that even if an explanation can mimic the predictions of the black 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. [84]. 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 differs 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. [3]. They fairwash a model by taking a black 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. [14]. The ability for developers to cover up unfairness in black-box models is one of the several vulnerabilities of explainable AI discussed in Hall et al. [26].

6. Humans as a Comparison Group for Explainable AI

Up to this point, we have outlined core concepts of explainable AI and related work in the field of computer science. However, an explainable AI system consists of both an AI system and a human recipient. To effectively understand both components, and to provide a benchmark for explainable AI systems, we next overview the explainability of human-produced judgments and decisions. Independent of AI, humans operating alone also make high stakes decisions with expectation that they be explainable. For example, physicians, judges, lawyers, and forensic scientists make decisions that can affect large populations. In these cases, a human makes the decision and provides their conclusion along with the evidence supporting that conclusion as an explanation. 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 [54]. 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.

6.1 Explanation

This principle requires only that the system 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 6.2, we will discuss whether human explanations are meaningful, and in Section 6.3, we will discuss the accuracy of those explanations.

Humans are able to produce a variety of explanation types [37, 53, 58]. However, producing verbal explanations can interfere with decision and reasoning processes [80, 81, 100]. 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 [17, 19, 44, 80]. 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., [53]).

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 [9, 88]. However, when judgments are provided via implicit categorization tasks, therefore bypassing an explicit judgment, lie detection accuracy can be improved [87, 88]. 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 black-box [53], and interfering with this black-box process can be deleterious to the accuracy of a decision.

6.2 Meaningful

To meet this principle, the system provides explanations that are intelligible and understandable. For this, we focused on the ability of humans to interpret how another human arrived at a conclusion. This concept can be defined operationally as: 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 is a gap between the ways forensic scientists report results and the understanding of those results by laypeople (see Edmond et al. [17], Jackson et al. [31] for reviews). Jackson et al. [31] 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 [31].

Complications for producing meaningful explanations for others include people expecting different explanation types, depending on the question at hand [58], context driving the formation of opinions [31], and individual differences in what is considered to be a satisfactory explanation [61]. Therefore, what is considered meaningful varies by context and across people.

6.3 Explanation Accuracy

This principle states that a system provides explanations which are faithful to the system's process for generating the output. 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 [62, 70, 99]. 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 [70]. People fabricate reasons for their decisions, even those thought to be immutable, such as personally held opinions [24, 34, 99]. 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 [95]. From a neuroscience perspective, neural markers of a decision can occur up to 10 seconds before a person's conscious awareness [85]. 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 [24, 25, 34]. 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 [75, 93].

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.

6.4 Knowledge Limits

This principle states that the system only operates under the conditions it was designed or that a provided output may not be reliable. 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 [41], 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 [7, 8, 12, 28, 63]. 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 [106]. 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 [103]. To use eyewitness testimony as a relevant example: although confidence and accuracy have repeatedly shown to be weakly related [86], 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 [101]. 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.

7. 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. 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. One example of adapting to meet practical requirements is illustrated by the trade-off between explanation detail and time constraints. These constraints highlight that certain scenarios require a brief, meaningful explanation to take priority over an accurate, detailed explanation. For example, emergency weather alerts need to be meaningful to the public but can lack an accurate explanation of how the system arrived at its conclusion. Other scenarios may require more detailed explanations but restrict meaningfulness to a specific user group; e.g., when auditing a model.

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 [23]. For humans, there are general tendencies for preferring simpler and more general explanations [58]. 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 [45]. Without conscious awareness, people incorporate irrelevant information into a variety of decisions such as first impressions, personality trait judgments, and jury decisions [21, 29, 90, 91]. Even when provided identical information, the context, a person's biases, and the way in which information is presented influences decisions [4, 15, 17, 23, 36, 43, 68, 94]. 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 [67]. 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 understanding 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 equips 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 Kristen Greene, Reva Schwartz, Brian Stanton, Amy Yates, and Jesse Zhang for their insightful comments and discussions.

References

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

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity Checks for Saliency Maps. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9505–9515. Curran Associates, Inc.

[3] Aivodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019). Fairwashing: the risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. ISSN: 1938-7228 Section: Machine Learning.

[4] Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg More Employable Than Lakisha and Jamal?: A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013.

[5] Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106(7):1039–1082.

[6] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M., and Eckersley, P. (2020). Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657.

[7] Bindemann, M., Attard, J., and Johnston, R. A. (2014). Perceived ability and actual recognition accuracy for unfamiliar and famous faces. *Cogent Psychology*, 1(1).

[8] Bobak, A. K., Mileva, V. R., and Hancock, P. J. (2018). Facing the facts: Naive participants have only moderate insight into their face recognition and face perception abilities. *Quarterly Journal of Experimental Psychology*, page 174702181877614.

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

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

[11] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA. Association for Computing Machinery. event-place: Sydney, NSW, Australia.

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

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

[14] Dimanov, B., Bhatt, U., Jamnik, M., and Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *European Conference on Artificial Intelligence*.

[15] Doleac, J. L. and Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal*, 123(572):F469–F492.

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

[17] Edmond, G., Towler, A., Growns, B., Ribeiro, G., Found, B., White, D., Ballantyne, K., Searston, R. A., Thompson, M. B., Tangen, J. M., Kemp, R. I., and Martire, K. (2017). Thinking forensics: Cognitive science for forensic practitioners. *Science and Justice*, 57(2):144–154.

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

[19] Fallshore, M. and Schooler, J. W. (1995). Verbal Vulnerability of Perceptual Expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,

21(6):1608–1623.

[20] Ferguson, T. (2014). *Game Theory*. Second edition.

[21] Flowe, H. D. and Humphries, J. E. (2011). An examination of criminal face bias in a random sample of police lineups. *Applied Cognitive Psychology*, 25(2):265–273.

[22] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). 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.

[23] Google LLC (2019). AI Explanations Whitepaper. pages 1–28.

[24] Hall, L., Johansson, P., and Strandberg, T. (2012). Lifting the Veil of Morality: Choice Blindness and Attitude Reversals on a Self-Transforming Survey. *PLoS ONE*, 7(9).

[25] Hall, L., Johansson, P., Tärning, B., Sikström, S., and Deutgen, T. (2010). Magic at the marketplace: Choice blindness for the taste of jam and the smell of tea. *Cognition*, 117(1):54–61.

[26] Hall, P., Gill, N., and Schmidt, N. (2019). Proposed guidelines for the responsible use of explainable machine learning.

[27] Haney, J. and Furman, S. (2019). Perceptions of Smart Home Privacy and Security Responsibility, Concerns, and Mitigations. *15th Symposium on Usable Privacy and Security*.

[28] Harvey, N. (1997). Confidence in judgment. *Trends in Cognitive Sciences*, 1(2):78–82.

[29] Hu, Y., Parde, C. J., Hill, M. Q., Mahmood, N., and O'Toole, A. J. (2018). First Impressions of Personality Traits From Body Shapes. *Psychological Science*, 29(12):1969–1983.

[30] IBM Research (Accessed July 8, 2020). Trusting AI. Available at https://www.research.ibm.com/artificial-intelligence/trusted-ai/.

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

[32] James, G., Witten, D., Hastie, T., and Tibshirani, R. (2014). *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, 1st edition in 2013, corrected 4th printing 2014 edition edition.

[33] Japkowicz, N. and Shah, M. (2014). *Evaluating Learning Algorithms A Classification Perspective*. Cambridge University Press.

[34] Johansson, P., Hall, L., Sikström, S., and Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, 310(5745):116–119.

[35] Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

[36] Kassin, S. M., Dror, I. E., and Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of Applied Research in Memory and Cognition*, 2(1):42–52.

[37] Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*,

57:227–254.

[38] Kim, B., Rudin, C., and Shah, J. A. (2014). The Bayesian Case Model: A Generative Approach for Case-Based Reasoning and Prototype Classification. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 1952–1960. Curran Associates, Inc.

[39] Koh, P. W. and Liang, P. (2017). 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. event-place: Sydney, NSW, Australia.

[40] Kroll, J. A., Huey, J., Barocas, S., Felton, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2017). Accountable Algorithms. *University of Pennsylvania Law Review*, pages 633–705.

[41] Kruger, J. and Dunning, D. (1999). 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.

[42] Kuhn, D. R., Kacker, R., Lei, Y., and Simos, D. E. (2020). Combinatorial Methods for Explainable AI. In *IWCT 2020*. Library Catalog: conf.researchr.org.

[43] Kukucka, J., Kassin, S. M., Zapf, P. A., and Dror, I. E. (2017). Cognitive Bias and Blindness: A Global Survey of Forensic Science Examiners. *Journal of Applied Research in Memory and Cognition*, 6(4):452–459.

[44] Kulatunga-Moruzi, C., Brooks, L. R., and Norman, G. R. (2004). Using comprehensive feature lists to bias medical diagnosis. *Journal of Experimental Psychology: Learning Memory and Cognition*, 30(3):563–572.

[45] Kveraga, K., Ghuman, A. S., and Bar, M. (2007). Top-down prediction in the cognitive brain. *Brain and cognition*, 65(2):145–168.

[46] Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-Velez, F. (2019). Human Evaluation of Models Built for Interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):59–67.

[47] Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016). Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1675–1684, New York, NY, USA. Association for Computing Machinery. event-place: San Francisco, California, USA.

[48] Lakkaraju, H. and Bastani, O. (2020). "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. Association for Computing Machinery.

[49] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2019). 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. Association for Computing Machinery. event-place: Honolulu, HI, USA.

[50] Lakkaraju, H. and Rudin, C. (2017). Learning Cost-Effective and Interpretable Treat-

ment Regimes. In *Artificial Intelligence and Statistics*, pages 166–175.

[51] Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. (2015). Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371.

[52] Li, O., Liu, H., Chen, C., and Rudin, C. (2018). Deep Learning for Case-Based Reasoning Through Prototypes: A Neural Network That Explains Its Predictions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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

[54] Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10(10):464–470.

[55] Luna, J. M., Gennatas, E. D., Ungar, L. H., Eaton, E., Diffenderfer, E. S., Jensen, S. T., Simone, C. B., Friedman, J. H., Solberg, T. D., and Valdes, G. (2019). Building more accurate decision trees with the additive tree. *Proceedings of the National Academy of Sciences*, 116(40):19887–19893.

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

[57] Marti, D. and Broniatowski, D. A. (2020). Does gist drive NASA experts' design decisions? *Systems Engineering*, (May 2019):1–20.

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

[59] Molnar, C. (2018). Interpretable Machine Learning.

[60] Molnar, C. (2019). *Interpretable Machine Learning*. @ChristophMolnar, online edition edition.

[61] Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI. *arXiv:1902.01876 [cs]*. arXiv: 1902.01876.

[62] Nisbett, R. E., Wilson, T. D., Kruger, M., Ross, L., Indeed, A., Bellows, N., Cartwright, D., Goldman, A., Gurwitz, S., Lemley, R., London, H., and Markus, H. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3).

[63] Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology*, 29(3):261–265.

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

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

[66] Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott,

C. L., and Sharpe, M. (2010). FRVT 2006 and ICE 2006 large-scale results. *IEEE Trans. PAMI*, 32(5):831–846.

[67] Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176.

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

[69] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2019). Manipulating and Measuring Model Interpretability. *arXiv:1802.07810 [cs]*. arXiv: 1802.07810.

[70] Pronin, E. (2009). The introspection illusion. In *Advances in experimental social psychology*, pages 1–67. Elsevier.

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

[72] Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory Valerie. *Judgment and Decision Making*, 7(3):332–359.

[73] Reyna, V. F. (2018). When Irrational Biases Are Smart: A Fuzzy-Trace Theory of Complex Decision Making. *Journal of Intelligence*, 6(2):29.

[74] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”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. ACM.

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

[76] Roach, J. (Accessed July 29, 2020). Microsoft responsible machine learning capabilities build trust in AI systems, developers say. Available at https://blogs.microsoft.com/ai/azure-responsible-machine-learning/.

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

[78] Rudin, C. and Radin, J. (2019). 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).

[79] Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C. S., Reynolds, D. A., Singer, E., Mason, L. P., and Hernandez-Cordero, J. (2017). The 2016 nist speaker recognition evaluation. In *Interspeech*, pages 1353–1357.

[80] Schooler, J. W. and Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71.

[81] Schooler, J. W., Ohlsson, S., and Brooks, K. (1993). Thoughts Beyond Words: When Language Overshadows Insight. *Journal of Experimental Psychology: General*,

122(2):166–183.

[82] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626.

[83] Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53.

[84] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). 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. Association for Computing Machinery.

[85] Soon, C. S., Brass, M., Heinze, H. J., and Haynes, J. D. (2008). Unconscious determinants of free decisions in the human brain. *Nature Neuroscience*, 11(5):543–545.

[86] Sporer, S. L., Penrod, S., Read, D., and Cutler, B. (1995). Choosing, Confidence, and Accuracy: A Meta-Analysis of the Confidence-Accuracy Relation in Eyewitness Identification Studies. *Psychological Bulletin*, 118(3):315–327.

[87] ten Brinke, L., Stimson, D., and Carney, D. R. (2014). Some Evidence for Unconscious Lie Detection. *Psychological Science*.

[88] ten Brinke, L., Vohs, K. D., and Carney, D. R. (2016). Can Ordinary People Detect Deception After All? *Trends in Cognitive Sciences*, 20(8):579–588.

[89] The Royal Society (2019). Explainable AI: the basics policy briefing. Available at https://royalsociety.org/-/media/policy/projects/explainable-ai/AI-and-interpretability-policy-briefing.pdf.

[90] Todorov, A. (2017). *Face value: The irresistible influence of first impressions*. Princeton University Press.

[91] Todorov, A., Mandisodza, A. N., Goren, A., and Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science (New York, N.Y.)*, 308(5728):1623–6.

[92] Toreini, E., Aitken, M., Coopamootoo, K., Elliot, K., Gonzalez-Zelaya, C., and van Moorsel, A. (2020). The relationship between trust in AI and trustworthy machine learning technologies. In *Conference on Fairness, Accountability, and Transparency (FAT* '20)*, Barcelona, Spain.

[93] Towler, A., White, D., and Kemp, R. I. (2017). Evaluating the feature comparison strategy for forensic face identification. *Journal of Experimental Psychology: Applied*, 23(1):47.

[94] Tversky, A. and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458.

[95] Tversky, A. and Shafir, E. (1992). The Disjunction Effect in Choice Under Uncertainty. *Psychological Science*, 3(5):305–309.

[96] Ustun, B., Spangher, A., and Liu, Y. (2019). Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 10–19, New York, NY, USA. Association for Computing Machinery.

event-place: Atlanta, GA, USA.

[97] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841.

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

[99] Wilson, T. D. and Bar-Anan, Y. (2008). The unseen mind. *Science*, 321(5892):1046–1047.

[100] Wilson, T. D. and Schooler, J. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology*, 60(2):181–192.

[101] Wixted, J. T., Mickes, L., and Fisher, R. P. (2018). Rethinking the Reliability of Eyewitness Memory. *Perspectives on Psychological Science*, 13(3):324–335.

[102] Woodruff, A., Fox, S. E., Rousso-Schindler, S., and Warshaw, J. (2018). A qualitative exploration of perceptions of algorithmic fairness. *Conference on Human Factors in Computing Systems - Proceedings*, 2018-April:1–14.

[103] Yeung, N. and Summerfield, C. (2012). Metacognition in human decision-making: Confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 367(1594):1310–1321.

[104] Zhao, Q. and Hastie, T. (2019). Causal Interpretations of Black-Box Models. *Journal of Business & Economic Statistics*, 0(0):1–10.

[105] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

[106] Zhou, X. and Jenkins, R. (2020). Dunning–Kruger effects in face perception. *Cognition*, 203(January).

[107] Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. (2017). Visualizing Deep Neural Network Decisions: Prediction Difference Analysis.