NISTIR 8367

Psychological Foundations of Explainability and Interpretability in Artificial Intelligence

David A. Broniatowski

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8367

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

NISTIR 8367

Psychological Foundations of Explainability and Interpretability in Artificial Intelligence

David A. Broniatowski
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8367

April 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 8367
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8367, 56 pages (April 2021)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8367

Abstract

In this paper, we make the case that interpretability and explainability are distinct requirements for machine learning systems. To make this case, we provide an overview of the literature in experimental psychology pertaining to interpretation (especially of numerical stimuli) and comprehension. We find that interpretation refers to the ability to *contextualize* a model's output in a manner that relates it to the system's designed functional purpose, and the goals, values, and preferences of end users. In contrast, explanation refers to the ability to accurately describe the mechanism, or implementation, that led to an algorithm's output, often so that the algorithm can be improved in some way. Beyond these definitions, our review shows that humans differ from one another in systematic ways, that affect the extent to which they prefer to make decisions based on detailed explanations versus less precise interpretations. These individual differences, such as personality traits and skills, are associated with their abilities to derive meaningful interpretations from precise explanations of model output. This implies that system output should be tailored to different types of users.

Key words

abstraction; implementation; fuzzy-trace theory

Table of Contents

List of Tables

List of Figures

1. Introduction

This paper draws upon literature in computer science, systems engineering, and experimental psychology to better define the concepts of *interpretability* and *explainability* for complex engineered systems. Our specific focus is on systems enabled by artificial intelligence and machine learning (AI/ML).

1.1 Why Define Interpretability and Explainability?

We focus on these terms because of their recent importance to the uptake of machine learning algorithms, as indicated by several recent laws that *require* algorithmic output to provide explanations or interpretations to users, who may differ significantly from one another in terms of their goals, education, or personality traits. For example, the Equal Credit Opportunity Act (ECOA)

> ...reflect[s] Congress's determination that consumers and businesses applying for credit should receive notice of the reasons a creditor took adverse action on the application or on an existing credit account...to help consumers and businesses by providing transparency to the credit underwriting process and protecting against potential credit discrimination by requiring creditors to *explain the reasons* adverse action was taken. [Ammermann] (emphasis added)

The implementation of ECOA in Regulation B further specifies that "A creditor must disclose the principal reasons for denying an application or taking other adverse action...and accurately describe the factors actually considered or scored by a creditor." [Com]

Additionally, the European Union's General Data Protection Regulation (GDPR) requires that AI systems provide human subjects about whom data are being gathered the right "...to obtain an explanation of the decision reached after such assessment and to challenge the decision." [111][1], with similar language reflected in France's *Loi pour une République numérique* [43] and ongoing debate about adopting similar regulations in the United States in the wake of California's adoption of the California Consumer Privacy Act of 2018 and its recently adopted amendment – the California Privacy Rights Act of November, 2020.

In parallel with attempts to address the societal concerns that drove the adoption of this legislation, major government investments (e.g., DARPA's eXplainable Artificial Intelligence [XAI] program) and highly-cited literature (e.g., [6, 42, 74, 87, 94, 125]) have attempted to define design requirements that engineers and computer scientists might adopt in order to determine whether their systems are interpretable or explainable. For example, Doshi-Velez and Kim [42] define model interpretability as a ML system possessing "the ability to explain or to present [output] in understandable terms to a human." Similarly, Singh defines an explanation by a ML model as a "collection of visual and/or interactive artifacts that provide a user with sufficient description of a model's behavior to accurately

[1] Recital 71, EU GDPR, https://www.privacy-regulation.eu/en/r71.htm

perform tasks like evaluation, trusting, predicting, or improving a model." (Singh, as cited in [61]). Gilpin et al. [54] posit that a good explanation occurs when modelers or consumers "can no longer keep asking why" in regards to some ML model behavior. Finally, Rudin [128] defines an interpretable machine learning model as one that is "*constrained in model form* so that it is either useful to someone, or obeys structural knowledge of the domain, such as monotonicity, causality, structural (generative) constraints, additivity, or physical constraints that come from domain knowledge." In contrast, she defines an explainable machine learning model as "a second (posthoc) model [that] is created to explain the first black box model".

Although these definitions identify interpretability and explainability as features of machine learning models, they point to important factors that are beyond the scope of traditional design: notions of simplicity, utility to the consumer, human comprehension, causal inference, interaction with domain knowledge, content and context, and social evaluation (such as trustworthiness).

These definitions, although localized to ML models, may be productively informed by decades of literature in experimental psychology, which treats *interpretability and explainability as psychological constructs*. The key insight of this literature is that *interpretation and explanation are distinct psychological processes*, characterized by distinct *mental representations*.[2] The question of whether or not a result is interpretable or explainable depends on the user. The designer must ask: "explainable or interpretable *for whom*?" (e.g. [136]).

1.2 Proposed Definitions of Interpretability and Explainability

Although the terms interpretability and explainability are frequently used interchangeably, especially in the computer science literature [6], this paper's fundamental argument is that *interpretability and explainability are distinct concepts.*

Interpretation refers to a human's ability to *make sense*, or derive meaning, from a given stimulus [71] (e.g., a machine learning model's output) so that the human can make a decision. Interpretations are simple, yet meaningful, *"gist"* mental representations that contextualize a stimulus and leverage a human's background knowledge (see Figure 1). A gist is a simple, yet productive, representation of a stimulus that nevertheless captures essential, or meaningful, distinctions that human users need to make informed, insightful decisions. Thus, *an interpretable model should provide users with a description of what a stimulus, such as a datapoint or model output, means in context*. In so doing, it enables that human to achieve insight by cuing values, goals, and principles which, in turn, enable high-level decision making.

Whereas humans rely on simple, imprecise gists to make decisions, machine learning models rely on programmatic *verbatim* processes to generate predictions. Explanations are relatively detailed mental representations that seek to describe the mechanisms underly-

[2] A mental representation of a stimulus (e.g., data or model output) is a symbolic image of that stimulus in a human's mind.

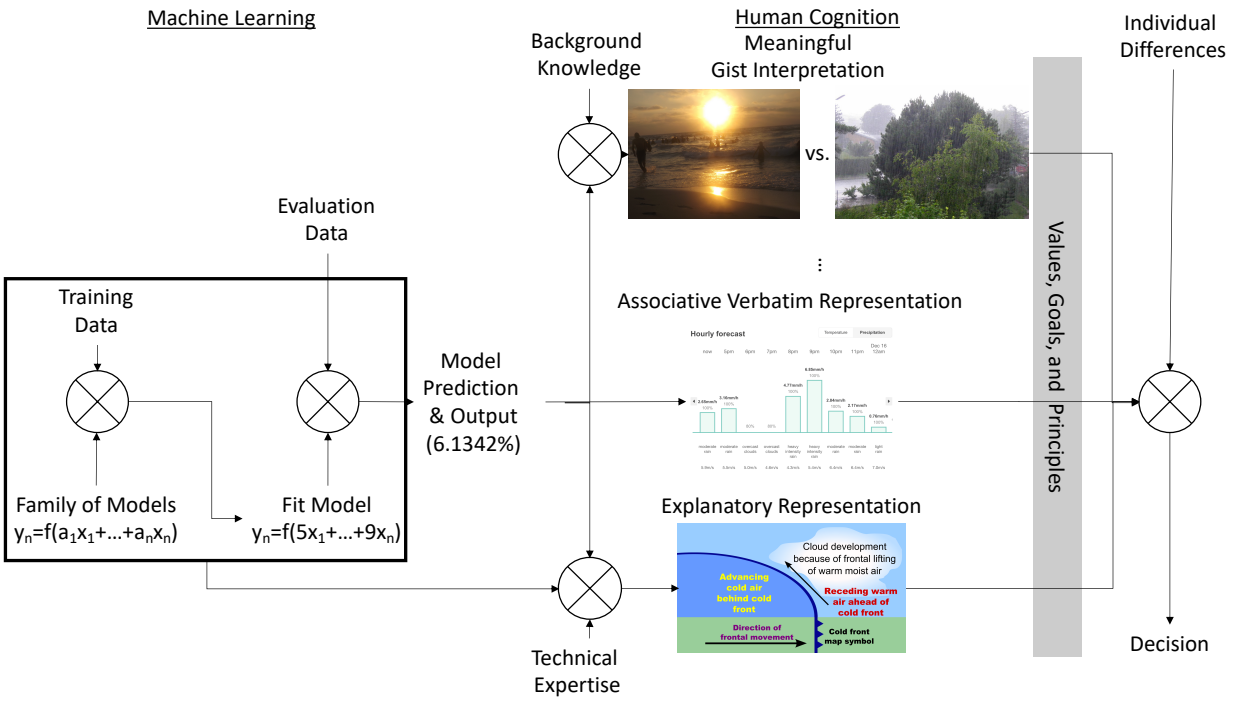


Fig. 1. Diagram of the interaction between a machine learning model and human cognition. For example, the initial training dataset may contain records of the hourly rainfall at a nearby beach. This trained model is then used to generate predictions from new evaluation data, such as a probability distribution over the amount of rain that a beachgoer might expect in a given hour. These predictions and other model output are then provided to the human as a stimulus. The human encodes the stimulus into multiple mental representations. The verbatim representation, is a detailed symbolic representation of the stimulus such as a graphical representation of the probability distribution over rainfall amounts per hour. In parallel, humans employ their background knowledge to encode a meaningful gist interpretation from the stimulus. For example, a simple categorical gist might be the distinction between "essentially no chance of rain" vs. "some chance of rain". Additionally, humans with appropriate expertise might be able to examine the form of the model to determine how it arrived at its conclusion. For example, a meteorologist with domain expertise might be able to examine the coefficients of a model's time series equations and recognize it as an indication of an incoming cold front. A human would then make a decision (e.g., whether or not to go to the beach) based on a combination of these representations. For example, a human without technical expertise might look at the stimulus and determine that the probability of rain is essentially nil, leading them to go to the beach (since the beach with no rain is fun, and the beach with some rain is not fun, and having fun is good). On the other hand, a human with meteorology expertise and data science expertise might recognize the signs of an oncoming cold front and realize that rain is a non-negligible possibility, leading them to choose another activity.

ing these verbatim processes: *an explanation of a model result is a description of how a model's outcome came to be*. Explanations thus seek to describe the process, or rules, that were implemented to achieve an outcome independent of context. Typically, explanations are detailed, technical, and may be causative. For example, an explanation may be a procedure outlining *how* a model achieved its result. Thus, explanations are typically more appropriate for technical practitioners, who can rely on extensive background knowledge to enable debugging tasks.

Although they are not necessarily verbatim processes themselves, explanations are, in this way, closer to verbatim mental representations than interpretations are. Whereas an interpretation seeks to make sense of a stimulus presented to a human subject, an explanation seeks to describe the process that generated an output.[3] Thus, an explanation of an algorithm's output is justified relative to an implementation, or technical process, that was used to generate a specific output. In contrast, an interpretation is justified relative to the functional purpose of the algorithm.

Explanations of ML algorithms can provide implementation details for how the algorithm carries out a known set of requirements. In contrast, interpretations justify these implementations in terms of the system's functional purpose. For example, the purpose of a Support Vector Machine classifier is to map datapoints into discrete classes, a task that must be justified in terms of the utility of the classification to a human decision-maker such as if this classifier is used to assign job applicants' resumes into merit-based categories during an interview process. The quality of the classification would then be evaluated relative to the requirements of this interview process – a classifier that is biased (e.g., that makes classifications based upon categories that are not merit-based, such as age, race, ethnicity, etc.) or that has a high error rate would be considered a poor classifier because it does not meet its functional purpose. In contrast, the explanation for why a particular classification decision was made is typically justified relative to its implementation. For example, when asking how a particular job candidate was classified as "not eligible", one must seek an explanation in terms of the algorithm's details, such as that the algorithm selected a set of candidates' profiles as "minimally acceptable" – i.e., they were support vectors – based on training data, and that this particular candidate's qualifications were, on the whole, inferior to those reference candidates. An even more detailed explanation would entail examining specific mathematical parameter values, such as the algorithm's regularization weights, to understand how specific attributes were combined and how support vectors were chosen.

In this paper, we make the case that explanations and interpretations are distinct mental representations that are encoded simultaneously, and in parallel, in the minds of the system's users. Furthermore, users differ from one another in the degree to which they are willing and able to utilize their own background knowledge to interpret detailed technical information. In effect, interpretable systems should provide no more detail than necessary to make a consequential decision, with the information provided justified relative to sys-

[3] Consider a car with a "check engine" light that is illuminated. An explanation might indicate that the check engine light turned on because the car's internal programming detected fuel flow irregularities. However, the interpretation for the driver is that the car needs to be taken to a mechanic for further evaluation.

tem's functional purpose. In contrast, explainable systems provide detailed mechanisms underlying how a certain implementation generated a certain output, regardless of what that output means to the decision-maker. An explanation seeks to replicate the decision in a more detailed manner, whereas an interpretation seeks to communicate the bottom line meaning.

The above definitions suggest that the efficacy of interpretations and explanations may differ between individuals, and indeed, we will review literature showing that they do so in systematic ways. That is, the audiences for these different types of outputs are likely to differ, such that developers who lack domain knowledge would be able to use a detailed mechanistic explanation to ensure that their design meets a specific functional requirement (e.g., a certain accuracy target), but may not understand the implications of this requirement for human users. In contrast, users who lack machine learning expertise but possess domain knowledge would likely find these detailed mechanistic explanations confusing, instead preferring a simple description of model output in terms of constructs with which they are familiar. Finally, a developer with domain knowledge can often bring this combined expertise to bear to make sense of a detailed mechanistic explanation in terms of its ultimate use case, thus ensuring that the algorithm moves beyond the rote requirements to best address the user's needs.

Disentangling explainability – whether one can describe the *mechanistic description of how a system made a specific prediction* – from interpretability – whether a human can *derive meaning from a system's output for a specific use case* – may form the basis for robust and reliable standards for explainable and interpretable ML system design, and should allow the development of standards that isolate technical design features from specific system functional requirements. This, in turn, should allow developers to segment the design process, such that system requirements may be defined at the appropriate level of abstraction. Additionally, we expect that better definitions of these terms will allow the ultimate development of metrics to assure compliance with these standards, thus enabling the creation of coherent regulatory policies for artificial intelligence that promote innovation while building public trust.

1.2.1 Illustrative Example: Rental Applications

Applications of machine learning for property rental have recently received negative attention given concerns about potentially discriminatory incidents[4] and potential violations of data privacy[5]. Under these circumstances, algorithmic interpretability may be an enabler of transparency, helping users better understand why a given decision was made.

[4]For example, the Landlord Tech Watch website reports on the technologies and data sources that landlords may use to deny applicants access to affordable housing: https://antievictionmappingproject.github.io/landlordtech

[5]For example, the "Locked Out" series reports on incidents of loan and rental applications that were denied because of improper data cleaning, recording, transportation, etc., and the difficulties that some renters applicants face in correcting these errors and, consequently, their personal records: https://themarkup.org/series/locked-out

For example, consider an algorithm that recommends rejection of a rental applicant. The algorithm would make this determination based on a family of mathematical models fit to training data, followed by evaluation of a model output generated from an additional held-out datapoint that represents the applicant's case. An interpretation of the algorithm's recommendation would contextualize the datapoint representing the applicant. A human would use their background knowledge to generate this context. For example, a human assessor might conclude that the applicant was a risk, based on the absence of the applicant's rental history. In contrast, a machine learning model would use a combination of training data and the model selected by the machine learning algorithm (including any associated sources of bias). Here, the algorithm might associate length of rental history with success, and therefore also categorize the applicant with a short history represents a financial risk. As will be discussed below, human interpretations differ from algorithmic ones in that the former are flexible whereas the latter tend to be brittle. Importantly, both interpretations are justified relative to a higher-level construct – a "rental history" – that contextualizes the decision relative to domain knowledge. Furthermore, this output provides the user with actionable *insight*. The solution to the problem is not to change the algorithm's implementation, but rather for the applicant to establish a rental history. In order to understand the meaning of this output, the applicant does not need to have any experience in AI or ML; rather, they must possess sufficient domain expertise to understand why rental histories are important indicators of approvals (we will discuss how interpretability may vary with domain expertise in section 2.3).

In contrast, an explanation of the same algorithm's output would start with the observation that the applicant was rejected and then seek to answer the question of how that decision came to be. For example, the explanation might specify that the algorithm was trained using a logistic regression classifier with specific coefficient values. Given the applicant's datapoint, one could then plug the values into the logistic regression equation, generate the model's probability of success for the applicant, and then observe that it is below the decision threshold. This explanation would not necessarily highlight the specific role of rental history, but a human analyst with access to this equation, and the expertise to interpret it, might observe that the largest marginal contribution to the algorithm's decision is the rental history.[6] Similarly, a human, asked to explain a rejection, might provide a causal explanation ("Your application was rejected because you don't have a rental history. People without rental histories are higher risk because they don't have any experience with paying rent on time, and because we don't have any evidence that they are responsible. As a rule, we prefer to rent to people with a reliable record of payments"). However, as we will discuss below, humans, and especially subject-matter experts, routinely violate such causal rules when making judgments. Arguably, this is because they are able to recognize necessary exceptions through the application of educated intuition (however, these same processes may also be a source of systematic bias if the underlying intuition is uneducated or otherwise inapplicable).

[6]Importantly, many modern machine learning algorithms, and especially deep neural networks, are not as easily analyzed by humans.

1.2.2 Illustrative Example: Medical Diagnosis

Like rental applications, medical diagnosis is a field about which concerns regarding algorithmic discrimination have been levied, necessitating transparency and, consequently interpretable AI.[7] Consider an AI system designed to make recommendations on antibiotic prescription for upper respiratory tract infections. For simplicity, we will again assume that this model is implemented using a logistic regression classifier with two classes corresponding to recommendations for, and against, antibiotic prescription. Finally, given the data reported to the system, suppose that the model has determined that the probability that a patient has a bacterial illness is 5%-10% [Cen]. The system would then provide the prescribing physician with a stimulus – the recommendation not to prescribe.

An *explanation* of this recommendation would reference the model's implementation. For example, the system might list the coefficients of the logistic regression model and the values of all of the model's variables (whether the patient has a sore throat, pain when swallowing, fever, red and swollen tonsils with white patches or streaks of pus, tiny red spots on the roof of the mouth, swollen lymph nodes in the front of the neck, cough, runny nose, hoarseness, or conjunctivitis [Wor]). Given these coefficients, the system might further explain that, when one multiplies the coefficients by the variable values, and then sums the result, the combined probability that the illness is bacterial is 5%-10%, indicating "No further testing nor antibiotics" [Wor]. This is where the explanation would stop.[8]

In contrast, an *interpretation* of the system's recommendation would reference simple, categorical representations of the relative risk and then link these to values. For example, the following values could apply: when the patient is sick getting better is good, whereas the staying sick is bad. Additionally uncomfortable side effects are bad (they will make the patient feel worse) and no side effects are good. Finally, unnecessary prescription promotes antibiotic resistance, potentially harming others (bad), whereas not prescribing has no effect on others. Given these values, the system would state: 1) the likelihood antibiotics would help is virtually nil; 2) antibiotics, if prescribed, could lead to uncomfortable side effects; 3) using antibiotics when they are not necessary could harm others [23, 24, 72], indicating "No further testing nor antibiotics".

Despite these recommendations, there are several reasons why an expert physician might prescribe antibiotics under these circumstances. For example, the expert might recognize that a patient is especially susceptible to bacterial infection, or might simply make the gist-based strategic choice that "it's better to be safe than sorry" [23].

[7] For example, a recent report [132] indicates how an attempt to adjust a statistical model to account for racial disparities in training data may have led to under-treatment for African-American patients needing kidney transplants.

[8] In practice an expert physician would probably not rely on rote output from this system, instead perhaps acknowledging that patient followup is needed if the symptoms don't clear up, since there is a possibility that the illness could be bacterial. Depending on the specific circumstances and context, this possibility could even lead the physician to prescribe "just in case" [24].

1.3 Historical context

Although interest in explainable artificial intelligence (AI) dates back to the development of expert systems in the 1980s [33, 56], explainability has recently reemerged as a desideratum for modern complex AI/ML systems. This is, in large part, due to the proliferation of such systems throughout society and because of the increasingly complex, and computationally intensive algorithms – sometimes trained on terabytes of data – that are being deployed to solve real-world problems.

This development is not unique to AI; rather it is the consequence of an increasingly complex infusion of technology into all aspects of society. Although our focus here is restricted to computational, and especially machine learning, technologies, these developments are part of a larger trend that extends throughout all areas of technology. [9] Pervasive embedded computation has accelerated this trend. It is rare to find a piece of technology that doesn't have some kind of computational component – from learning thermostats to credit score adjudications to visa applications. These technologies also require several different types of expertise to regulate appropriately. First, technical expertise is required to understand how these technologies operate and, because the technologies are so complex, this expertise is restricted to a relatively small number of people, whereas the number of people whose lives are directly impacted has grown significantly. However, many types of expertise are relevant. For example, evaluating the legal consequences of AI/ML technologies requires in-depth familiarity with relevant areas of law. Similar concerns may apply to financial credit assessment, job application review, issues of political and social equity, and other ethical concerns. Thus, it is not sufficient to query the experts in a specific area. Effectively assessing, and thus regulating, interpretable and explainable AI systems requires pooling expertise from across fields that have not traditionally interacted.

We can expect the pace of this trend towards increasingly complex systems to accelerate. This evolution in major technological trends has been documented by the Engineering Systems movement. [39] Started in the early 2000s, this movement recognized that technology and modern society are highly intertwined and that the pace of technological and social change requires that the design of complex systems must adapt to account for what these scholars have called the "ilities".[10] Explainability and interpretability are both "ilities" and exhibit similar difficulties associated with their measurement. "Ilities" have historically been subject to problems of both polysemy, meaning that the same terms are frequently used to describe distinct concepts, and synonymy, meaning that different terms

[9] Consider automotive technologies: A mechanically-inclined person in the 1950s could often diagnose and fix problems with their own automobiles. However, as these automotive systems increased in complexity, professional expertise became increasingly necessary. By the early 2000s, with the rise of embedded computing, specialized tools became necessary to even diagnose, let alone fix, problems. For example, a car's internal computer may trigger a check-engine light, but provide no additional information to the driver. Further diagnosing this problem instead requires a specialized tool, which provides a diagnostic code that must be interpreted by trained mechanics who specialize in that particular brand of automobile.

[10] So called because several of them end with the suffix "ility". Key "ilities" include flexibility – the need for a system to adapt to changes in its environment – survivability – the need for a system to continue operation despite disruption, etc.

sometimes refer to the same underlying construct [28]. Furthermore, these terms entail a significant social component that cannot be disentangled from core values of users, designers, and decision-makers. Finally, "ilities" have a strong policy component because they cannot be studied in isolation from their effects on the public, and especially on vulnerable populations. Thus, attempts to define explainability and interpretability in artificial intelligence are comparable to challenges faced by scholars studying other complex engineered systems [39], for which definitions of abstract, yet important, concepts such as flexibility, resilience, etc., strongly depend on social evaluations. Two decades of research in this area have discovered that these highly-abstract requirements may be difficult to measure in a standard way because of their highly context-sensitive, socially-contingent nature. Nevertheless, their importance justifies the challenge of establishing standards that are flexible enough to be responsive in different contexts.

2. The Psychology of Interpretability and Explainability

The definitions of interpretability and explainability proposed in this paper build upon decades of empirical research in experimental psychology (see [115, 118] for reviews). We draw upon this extensive literature to posit a distinction between interpretation – the process of deriving meaning from a stimulus (such as a model's output) – and explanation – the process of generating a detailed description of how an outcome was reached (see [55] for preliminary data supporting this claim). We argue that the relationship between human decision making and algorithmic decision making is analogous to different levels of *mental representation*. Human individual differences also consistently predict *which* human subjects prefer to rely on these different representations when making consequential decisions, especially about how to use numerical information [27].[11]

2.1 Interpretations Provide Meaning in Context

Interpretable machine learning is concerned with helping humans generate interpretations of data and model output. Thus, we review literature in human psychology pertaining to how humans derive interpretations from stimuli, and especially quantitative stimuli. A leading theory in this area, Fuzzy-Trace Theory, posits that humans encode stimuli into multiple mental representations simultaneously and in parallel [115] (see Figure 2). These mental representations vary from one another in their level of precision, with humans preferring to rely on the least precise representation that still makes a meaningful distinction when making a decision.

[11] For example, a recent preprint provides some preliminary data suggesting that one leading algorithm for explainable AI [126] did not improve decision-makers' objective *accuracy* – i.e., their ability to retrieve the right (verbatim) number – whereas a more accurate AI system did. Similarly, a recent human subjects experiment [107] shows that model transparency – i.e., a model with a small number of features – did not improve humans' prediction accuracy and may have actually impaired their ability to correct inaccurate predictions due to information overload (see also [55]).

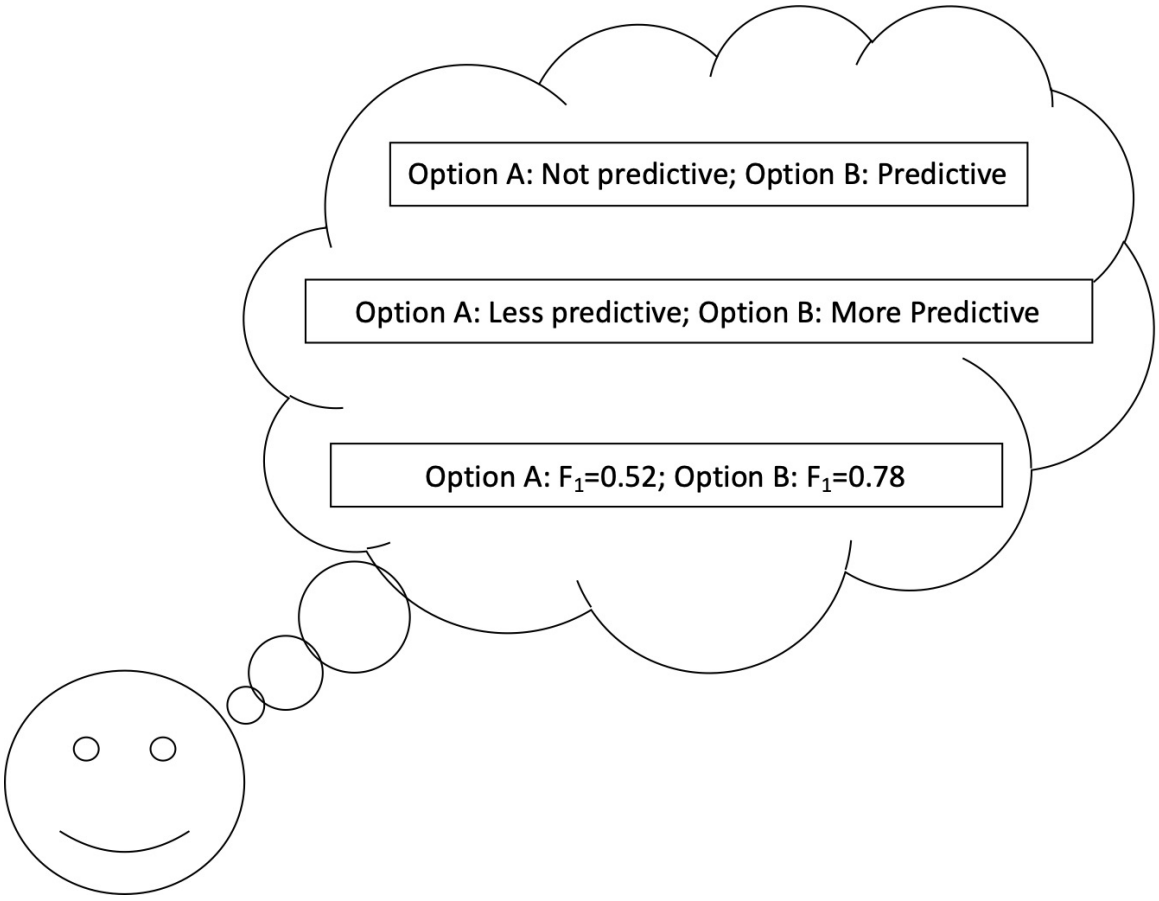


Fig. 2. Multiple Levels of Mental Representation Encoded in Parallel.

Table 1. Performance of three hypothetical models to detect online malicious behaviors.

Model	Accuracy	Precision	Recall	F_1 Score
Naive Bayes	0.267	0.563	0.897	0.692
Support Vector Machine	0.732	1.000	0.716	0.834
k-Nearest-Neighbor	0.524	0.834	0.637	0.722
Logistic Regression	0.907	0.859	0.617	0.718

Humans tend to make decisions based on the simplest of these representations – the *gist* interpretations of stimuli. Humans may encode multiple gists at varying levels of precision, forming a hierarchy of gists [27]. In contrast, algorithms follow only rote *verbatim* processes when making predictions. Humans also encode verbatim mental representations, which are simply detailed representations of the stimulus itself (e.g., raw system output). These different levels of mental representation of a stimulus are encoded simultaneously and in parallel [115]. Furthermore, these representations may compete with, or build upon one another [22] when providing input into human decision-making.

2.1.1 Categorical Gists

Mental representations are hierarchical in nature, with humans preferring to make decisions on the least precise, often categorical, representation of a stimulus. These categories nevertheless make *meaningful* distinctions. For example, for numbers, these categorical representations often take the form of simple contrasts such as between "some" and "none" of a quantity. (The categorical distinction between "some" and "none" is one of the most basic gists [22].) Under these circumstances, humans draw upon their prior knowledge when making these determinations. For example, consider a set of machine learning models that are designed to detect malicious online behavior. A social media platform might use this classifier to automatically remove accounts that seem to violate the platform's Terms of Service (see Table 1).

A computer scientist evaluating these classifiers may notice that the k-Nearest-Neighbor (kNN) classifier has an accuracy of 52.4% on this binary classification task and determine that it has "essentially no" predictive accuracy (where 50% is the performance of a random coin flip). Notably, this assessment requires some background knowledge: 1) that there are only two classes; 2) that the classes are balanced in the training set. In contrast, the accuracy of the other two models would both be an improvement over that of the SVM. Both would have "some" accuracy. In this case, it would rule out the kNN classifier. However, these gists are not just simple – they are also *insightful*. For example, the Naive Bayes Classifier has an accuracy of 26.7% which, although the smallest value, should also be classified as having a gist of "some accuracy" since a computer scientist would recognize that, for binary classifiers, an accuracy of 26.7% is equivalent to an accuracy of 73.3% if one simply flips the class labels. In contrast, a novice applying verbatim rules would not share this insight and might erroneously consider the Naive Bayes Classifier to be less predictive

than the kNN classifier.

Humans, when deciding, must draw upon *values* [135] to determine which category, in a binary pair, is better. Here, some accuracy is good and no accuracy is bad. These binary valences are held in long-term memory [114] and constitute part of what the human brings to the evaluation process.

2.1.2 Ordinal Gists

Several options can have the same gist. For example, all but the kNN classifier models have "some" accuracy. Here, categorical gists do not make meaningful distinctions and therefore cannot assist a decision. To distinguish between these classifiers, a more precise level of mental representation can be used. The Logistic Regression classifier has "more accuracy" than the other classifiers. "More" vs. "less" is an ordinal gist. However, this gist is only helpful in selecting a model when there is a single evaluation metric. In practice, ML models may be evaluated using several metrics. For example, the Naive Bayes model has higher recall, but lower precision, whereas the Logistic Regression model has higher precision but lower recall. Thus, these models cannot be ranked along these dimensions using only an ordinal gist.

2.1.3 Precise Verbatim Representations

Typically, practitioners try to reduce these multiple metrics to a single metric for comparison purposes. The process of deriving these composite, precise metrics requires rote application of mathematical rules. For example, one may rely on a composite metric, such as the F_1-score (i.e., the harmonic mean of precision and recall). In this example, the Logistic Regression and Naive Bayes classifiers have equal values of F_1, which means humans (or algorithms) relying on this verbatim rule would be indifferent between the two. In contrast, the Support Vector Machine has the highest F_1 score of the three models.

2.1.4 Moving Beyond Rote Optimization: Gists are Insightful

Does this mean that the Support Vector Machine is the best model? Although humans encode multiple levels of representation in parallel, we prefer to make decisions based on gists whenever possible. These gists are not arbitrary, but correspond to meaningful distinctions. In the case of machine learning models, we use these models to achieve a goal. In the example above, the classifier was used to identify malicious online behavior. In the context of this task, it makes sense to favor precision over recall and accuracy since the consequences of a false positive are significantly worse than those of a false negative. A novice might blindly apply this rule and thus select the Support Vector Machine. However, a human expert would not necessarily do so. Consider that the Support Vector Machine has a precision of 1.000 – a "perfect score". Although someone relying on the verbatim representation – such as an inexperienced student – might judge this to be the best possible precision score, an experienced modeler would recognize that such a high value could be an

indicator of a problem in the algorithm's implementation. For example, one might achieve perfect precision if only a very small number of cases are being correctly classified. The corresponding gist would be "too good to be true". As described above, the kNN classifier would be ruled out because it has a gist of no accuracy which, regardless of precision, is problematic. Thus, the human expert might rely on an ordinal gist to choose the model with "better" precision – the Logistic Regression model – because contextual cues indicate that the other two models are inferior for the functional purpose – i.e., the goal – of the machine learning task. Research based on Fuzzy-Trace Theory has shown that models that emphasize the gist – such as by displaying output in ways that more readily allow users to draw meaningful conclusions – inspire greater trust, confidence, and understanding [37, 146**?**]. This implies a clear design goal for ML system designers who are concerned with interpretability – system output must communicate the gist.

2.2 Explanations Emphasize Implementation

One of the key tenets of Fuzzy-Trace Theory is that humans encode multiple mental representations. Whereas interpretations are mental representations that communicate new categorical insights, *explanations are detailed mental representations that communicate the implementation mechanisms that led to a particular output*.

2.2.1 Causal Mental Models

Several theories of explanation emphasize the importance of inferring the "causal chain" leading to a specific model's output (for a review, see [94, 95])[12]. Lombrozo's [86] review of explanations in human cognition also indicates that causal structures are one enabler of explainability [13] Following these traditions, the XAI literature has focused on causality as the theoretical foundation of explanation. For example Hoffman & Klein [65] relate explanation to causal inference, especially arguing that humans explain things by creating prospective (i.e., predictive) causal explanations and highlighting a lacuna in the literature around this specific type of explanation, instead arguing that most prior work has focused on physical causality and retrospective causality. Hoffman et al., [66] describe prior work extracting various different structures of causal chains pertaining to events around the world, and Klein [74], further develops this theme by claiming that causal networks can and should be the basis for communicating explanations. Finally, Hoffman et al. [64]

[12] Beyond the role of causality, these theories assert that explanations are "contrastive" (e.g., Why P rather than Q?). As will be discussed in 2.3 some human subjects are more likely to carry out these contrastive comparisons between successive stimuli. Beyond causal relationships, Miller also emphasizes that explanations also entail social attributions [88].

[13] One of Lombrozo's primary insights is that explanations are *not* exclusively causal, but rather that similarity and diversity of stimuli can help subjects derive general principles that go beyond the surface forms of the causal structures that one might infer. These general principles draw upon extensive prior knowledge that make certain explanations more probable than others. This implies that attempts to define "explainability" (here, understood as encompassing both explanation *and* interpretation) entirely based on causality (or conversely, on counterfactual reasoning) may be missing an essential element – generalizability/gist.

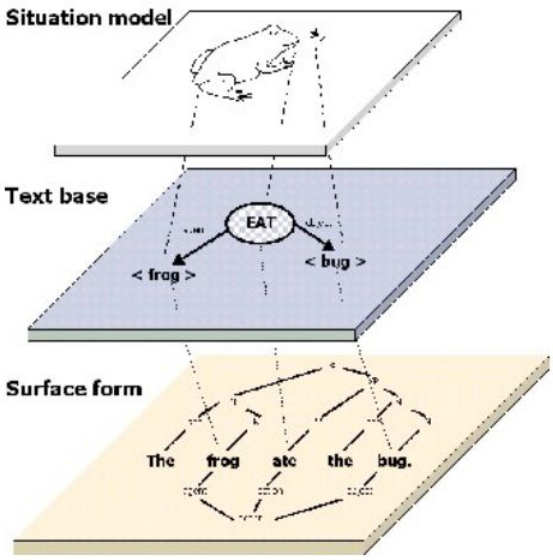


Fig. 3. Representational Hierarchy for a Text. By Original uploader was Aschoeke at en.wikibooks. Later version(s) were uploaded by Asarwary at en.wikibooks. - Transferred from en.wikibooks, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=5063569

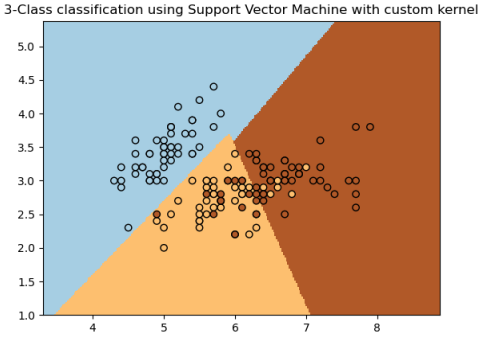
emphasize the role of exploration in forming causal explanations, differentiating between global and local explanations that accord with users' mental models, and emphasizing local explanations' needs for contrastive or counterfactual accounts; implicitly putting causal structures at the center of an explanation. According to these scholars, *an explanation is, at its core, a causal mental model*.[14]

These claims are supported by extensive prior work in psycholinguistics and narrative reasoning, especially in the legal domain. Several decades of prior work in psycholinguistics have emphasized the role of causal structure building in generating a "situation model" – i.e., a structured mental representation – of a given text (see Figure 3). By analogy, a similar hierarchy may be constructed for ML model output (see Figure 4).

Causal relations are among the most important (although certainly not the only) types of inferences that are extracted from narrative texts by readers seeking to comprehend a text. Furthermore, studies in psycholinguistics have identified a narrative's causal coherence as a key factor driving a story's comprehensibility [137, 143]. Although several dimensions of narrative coherence have been identified [50, 110, 151], there is a consensus within the literature that coherent narratives allow readers to construct causal situation models of the events described [41, 51, 89, 138, 143]. By analogy, one might expect that system

[14]Notably, these authors identify several structures beyond causal chains, including abstractions, "swarm", etc., and relate these to different types of explanations, concluding that "The property of 'being an explanation' is not a property of statements: It is an interaction of statements with knowledge, context, and intention."

Model Output



Mathematical Basis

$$\text{maximize } f(c_1 \dots c_n) = \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i (\varphi(\vec{x}_i) \cdot \varphi(\vec{x}_j)) y_j c_j$$

$$= \sum_{i=1}^{n} c_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i c_i k(\vec{x}_i, \vec{x}_j) y_j c_j$$

Surface Form (Code Base)

```
print(__doc__)

import numpy as np
import matplotlib.pyplot as plt
from sklearn import svm, datasets

# import some data to play with
iris = datasets.load_iris()
X = iris.data[:, :2]  # we only take the first two features. We could
                      # avoid this ugly slicing by using a two-dim dataset
Y = iris.target


def my_kernel(X, Y):
    """
    We create a custom kernel:

                (2  0)
    k(X, Y) = X  (   ) Y.T
                (0  1)
    """
    M = np.array([[2, 0], [0, 1.0]])
    return np.dot(np.dot(X, M), Y.T)


h = .02  # step size in the mesh

# we create an instance of SVM and fit out data.
clf = svm.SVC(kernel=my_kernel)
clf.fit(X, Y)

# Plot the decision boundary. For that, we will assign a color to each
# point in the mesh [x_min, x_max]x[y_min, y_max].
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, h), np.arange(y_min, y_max, h))
Z = clf.predict(np.c_[xx.ravel(), yy.ravel()])

# Put the result into a color plot
Z = Z.reshape(xx.shape)
plt.pcolormesh(xx, yy, Z, cmap=plt.cm.Paired)

# Plot also the training points
plt.scatter(X[:, 0], X[:, 1], c=Y, cmap=plt.cm.Paired, edgecolors='k')
plt.title('3-Class classification using Support Vector Machine with custom'
          ' kernel')
plt.axis('tight')
plt.show()
```

Fig. 4. Representational Hierarchy for a Support Vector Machine.

users seeking to understand or explain model output would be assisted by coherent causal explanations that would explain how a system came to a certain conclusion.[15]

This interpretation of the literature is supported by an extensive body of work on mental models, which studies how technical experts represent, and make decisions about, complex systems. Although a full review of this literature is beyond the scope of this paper, scholars agree that mental models "represent (perceived) cause-and-effect dynamics of a phenomenon" [68], thus helping people to make predictions. Thus, literature that seeks to derive causal descriptions of complex system operation is concordant with classical mental model theories. Finally, these causal representations are also widely used in legal reasoning. For example, the "story model" of juror decisionmaking [103] generally assumes that jurors jointly construct a causal story regarding the facts of a court case, and that these causal stories are then matched to verdicts. Similar techniques for building "cognitive maps" have been applied to fields as diverse as political science [10], and power plant design [99]. Thus, several domains have independently converged on the same conclusion – that comprehension of mechanisms is facilitated by a structured causal model. However, these domains also agree that applying these mechanisms to real-world problems must go beyond causal reasoning. In the story model, jurors must match story structure to knowledge about verdicts. In the mental model literature, Rasmussen's [108] abstraction hierarchy has been widely applied to demonstrate the contingency of causal representations on functional purpose. Finally, in the narrative comprehension literature, causal structures exist at multiple levels, with several substructures. Furthermore, the most meaningful levels in these substructures interface with other narrative elements associated with preferences (e.g., goals and characters) [58]. Importantly, in each of these cases, later research has documented that causal explanations are *interpreted* before a decision is made. Beyond Rasmussen's abstraction hierarchy, which applies to causal mental models [99], scholars of legal reasoning have found that amounts awarded as damages for legal verdicts depend on categorical and ordinal contextual cues that allow jurors to compare amounts to a meaningful reference point [63].

Thus, although one might think that a meaningful interpretation is a consequence of an unambiguous or otherwise precise explanation, Fuzzy-trace theorists have shown that different mental representations are encoded in parallel. This means that a mental representation that provides an interpretation can be distinct from a mental representation providing an explanation, and vice versa. As will be discussed in the next sections, the choice of mental representation on which to rely is also a function of individual differences in skills and personality traits.

2.3 Individual differences

Although one might think that a meaningful interpretation is a consequence of an unambiguous or otherwise precise explanation, Fuzzy-trace theorists have shown that different

[15]These structures, in turn, can facilitate extraction of gist of the story (see [25], for a review); however, the causal structures are not, themselves, the gist.

mental representations are encoded in parallel. Meaningful interpretations and mechanistic explanations are often *not* derived from one another or from precise verbatim data. The choice of mental representation on which a human relies is also a function of individual differences in skills and personality traits.

Humans differ from one another in systematic ways. Some of these differences are matters of *skill*. For example, a professional computer scientist with years of training is endowed with a skillset that is quite different from that of a professional legal scholar. Individuals with relevant skills may prefer to rely on more precise levels of mental representation if they have the ability to process them. For example "numeracy" – mathematical ability [44, 84, 104] – allows individuals to make sense of complex numeric data such as percentages and fractions, such that they are less susceptible to statistical bias when making decisions. Similarly, in a machine learning context, [67] found that users with computer science backgrounds (and especially doctoral-level training) were more likely to agree that the system was useful and trustworthy if they understood *how* the system worked (and vice versa), and Linderholm et al. [85] found that more skilled readers, and those with more relevant background knowledge, were better able to extract the gist from narratives with poorly-defined causal structures. These interpretive processes are associated with domain expertise [115] – a hallmark of gist processing [121].

Other differences are matters of personality *traits*. For example, some individuals prefer to rely on their "gut feelings" – i.e., their intuitive judgments – when making a decision, whereas others prefer to engage in extensive deliberation. The Cognitive Reflection Test (CRT; [46]) is a measure of this trait (although it is also correlated with numeracy and intelligence [84, 104]), and researchers have found that individuals with high CRT are less susceptible to decision biases that oppose intuitive to deliberative modes of thought (such as the well-known "framing effect"; [141]). Similarly, the Need for Cognition (NFC) Scale [31, 32] measures subjects' preferences to exert mental effort. For example, [57] describes evidence for a model of narrative comprehension in which multiple levels of mental representation are encoded, with some readers preferring to use coherence-building strategies relying on effortful "close-to-the-text" reading and those who utilize a more interpretive strategy that is "farther" from the text. In the domain of decision-making under risky, researchers have found that individuals possessing high NFC scores are more likely to answer several risky choice framing problems consistently [38, 82], presumably because they exert effort to notice similarities or contradictions between different problems with similar structure. This explanation of these findings is supported by evidence that within-subjects comparisons between stimuli can lead subjects to censor gist-based responses when contradictions are detected, thus encouraging subjects to focus on more detailed features [27]. Similarly, research has shown that some human subjects have difficulty making determinations about whether models are "fair" or "just" – both categorical gists – in the abstract (i.e., absent important context), and instead compare those explanations to prior experience or to a second system, enabling ordinal comparisons ("more fair/just" vs. "less fair/just") [12]. For this reason, Mittelstadt and colleagues [97] argue that models should be contrastive to facilitate interpretability. However, these authors also take pains to emphasize that such

contrastive explanations frequently miss important context – i.e., they spur reliance on decontextualized verbatim representations.

The above discussion implies that there is no single measure for interpretability or explainability that applies to all humans; however, there may be a measure that can be defined relative to the expected distribution of skills and personality traits for each intended audience. Future work should therefore focus on characterizing these factors within user communities.

2.3.1 Experts prefer to rely on meaningful interpretations

Above, we stated that most individuals reason, recall, and prefer to rely on less precise representations when making decisions [118]. This reliance on gist representation is a developmental feature of human cognition: compared to non-experts, experts are more likely to rely on gist representations in their domains of expertise [7, 123, 124]. Fuzzy Trace Theory [115] therefore distinguishes between rote knowledge – recall of verbatim facts or associations – and insightful expertise. Compared to novices, experts are better able extract the *essence*, or most relevant information, and ignore less meaningful details [123]. Experts have therefore developed intuitive categorical representations of stimuli that are simple, yet powerful, and enable them to make decisions. For example, NASA engineers rely on categorical determinations of "costly" or "costless" when making determinations about launching cargo missions [90], whereas expert physicians rely on categorical determinations of risk – reflecting an intelligent strategic choice that takes the very low probability, but non-negligible possibility, that the patient may require antibiotic therapy – when treating very sick patients who might require antibiotics [23, 24, 72].

2.4 Relationship of Fuzzy-Trace Theory to Prior Theories

Fuzzy-Trace Theory moves beyond alternative accounts that are found in the AI and psychology literatures.

2.4.1 Schema Theories and Association Theories

Several prior theories may be categorized into two broad groups: schema theories and association theories (e.g., [49, 133]). Schema theories posit that humans use higher-level data structures – called "schemata" or "frames" – that impose "top-down" structure on memories and experiences, making sense of world stimuli, and thus imposing biases. In contrast, association theories assume that meaning emerges "bottom-up" from frequently-observed patterns that co-occur in the world. Rather than making sense of these co-occurrence patterns, associationist theories posit that meaning is simply a function of statistical regularity. As early as 1983, Alba and Hasher [8] found that human memory displayed characteristics of both schematic and associationist theories. However, elements of both models were also repeatedly falsified, meaning that neither schematic nor associationist theories could account for all of the experimental findings (see [26, 115] for a more detailed exposition).

Fuzzy-Trace Theory explains these contradictory findings with the core theoretical distinction between gist and verbatim mental representations, that are encoded distinctly, yet in parallel (gist representations are *not* derived from verbatim representations). Although humans prefer to rely on gists, they also encode, and can therefore recognize, verbatim representations. In contrast, algorithms are verbatim by their very nature. Thus, humans working together with ML algorithms may get the best of both worlds – applying gist-based structured background knowledge to interpret association-based algorithmic output.

2.4.2 Heuristics and Biases

The key construct of Fuzzy-Trace Theory, gist, also moves beyond other theories that posit reliance on intuitive judgment. The "heuristics and biases" research paradigm (e.g., [53]) also acknowledges the role of intuition in human behavior, but considers intuitive judgments to be primitive and therefore associated with poor decision-making. This tradition points to routine human violations of statistical and decision-axioms as evidence for this claim; however, developmentally-advanced educated intuition often leads to *better* outcomes [112] even when experts may achieve these outcomes for what external observer may consider to be the "wrong reasons" (see 4.1). For example, an experienced physician may make the right decision regarding how to treat a patient given test results, even though their mathematical calculations regarding the numerical probabilities that they assign to different treatment outcomes may be incorrect [7]. Indeed, gist representations of complex problems allow experts to make decisions that are driven by context that is grounded in extensive background knowledge [17, 113, 117, 118, 122, 123]. This context allows experts to focus on the *essence* of information when making decisions, neglecting less important features that do not deliver insight [112]. Thus, gist representations, when informed by expertise, are not just based on a rote simplification, but are rather driven by *insightful* simplifications that are meaningful to decision-makers.

Evidence in favor of Fuzzy-Trace Theory's account of decision-making shows that the theory is both scientifically parsimonious and has more predictive accuracy than does Cumulative Prospect Theory [142] – the leading theoretical account in the heuristics and biases tradition – which nevertheless cannot account for key experimental effects that Fuzzy-Trace Theory does account for (see [22, 115] for details).

2.4.3 Naturalistic Decision Making

Naturalistic Decision Making [73, 150, KLEIN et al.], another leading framework that is especially popular in the human factors engineering and XAI literatures, posits that humans draw upon their prior experience to recognize patterns, which, in turn, drive decisions [76]. Both Naturalistic Decision Making and Fuzzy-Trace Theory acknowledge the role of intuition in improving decision-making; however, decisions that are based on gist intuitions are not simply "recognition primed decisions" as posited by the Naturalistic Decision Making tradition [76]. Rather, context cues (such as when subjects are encouraged to think about a problem from a medical or statistical perspective) can influence reliance on the level of

mental representation [15] meaning that recognition does not guarantee that a decision will rely on expert intuition. Whereas recognition is a rote verbatim strategy (theorized by associationism), gist representations bring background knowledge to bear, contextualizing scenarios such that they make sense and therefore providing insight to the human decision-maker. In fact, an extensive body of literature shows that humans can recognize both gist and verbatim representations in parallel, and yet prefer to rely on the gist when making decisions [18, 19, 115, 119].

Thus, an extensive body of literature supports the contention that Fuzzy-Trace Theory is both more parsimonious and more predictive than competing theoretical accounts of the role of interpretation in judgments and decisions. These findings apply to both texts, such as are found in the legal reasoning domain, and numerical stimuli such as are found in the engineering domain [90] or generated by machine learning models.

3. Computer Science Definitions of Interpretability and Explainability

The above discussion emphasizes that interpretability and explainability are functions of the user, the use case, and other contextual factors, as much as they are functions of the system being used. However, the psychometric properties of users are generally not under designers' control. Here, we discuss the state-of-the-art for explainable AI algorithms, and how systems might be designed to promote interpretability and explainability.

3.1 Comparison of Mental Representations to Current Machine Learning Paradigms

Whereas humans generate multiple mental representations in parallel, "shallow learning" algorithms generate a single model, or distribution of models from the same mathematical family, when representing a dataset – a verbatim process. Beyond shallow learning, several ML techniques do generate multiple representations. For example, ensemble learning is a process by which multiple models are generated and then ultimately aggregated to form a single hypothesis. However, these models do not differ from one another in terms of their level of precision – they simply apply different families of mathematical operators to the same set of features. In contrast, multitask learning algorithms seek to replicate the flexibility of human gist representations by training a model to generate a common representation of several stimuli from different domains, thus enabling "far transfer". When successful, these models may learn more abstract representations that are superficially similar to gist representations; however, they still generate only a single model. Finally, deep neural nets generate multiple representations of a dataset; however, they do so by deriving abstract representations from more concrete representations, whereas humans encode these representations *simultaneously* and *in parallel*, meaning that humans do *not* derive more simple interpretations from more detailed representations [118].

3.2 Algorithmic Paradigms Designed to Promote Interpretability and Explainability

A recent comprehensive literature review of computational approaches to explainable AI notes that, for computer scientists, the concepts of interpretability and explainability are "closely related" [6]. These authors assert that "intepretable systems are explainable if their operations can be understood by humans" (pp. 52140-52141; emphasis added). Although explainability and interpretability are sometimes used interchangeably in the computer science literature, this review provides data supporting the contention that "in [the] ML community the term 'interpretable' is more used than 'explainable"' (p. 52141; see also [42]), especially when compared to the usage of these terms by the general public. Consistent with the psychological definitions outlined above, this finding may indicate that producers of AI products are more able to interpret the output of these systems because they posses specialized background knowledge. Indeed, Bhatt et al. [11] posit that this distinction may belie a difference in the design goals of these user groups: algorithm developers generally seek explanations so that they may debug or otherwise improve their algorithms, and they might therefore develop explainable AI tools for that purpose. Thus, an explanation is generally understood by computer scientists to indicate *how* a computational system arrived at, or generated, a certain output. A good explanation is often *causal* and justified relative to a system's implementation – e.g., "the algorithm is biased towards visa application refusal BECAUSE the training data are unbalanced". This sort of explanation is quite useful for debugging these complex systems, but only if the user has the appropriate background knowledge and technical expertise to do so.[16] For example, the explanation given above would lead a developer to gather more balanced data and retrain the algorithm, but would not suggest an immediate action to an end user, except perhaps to abandon use of the algorithm.

3.2.1 Local Feature Importance

Much of the work in explainable artificial intelligence attempts to help developers to determine simple rote verbatim associations between inputs and outputs, with the aim of assisting them to infer potential causal mechanisms. For example, the local feature importance paradigm (e.g., [33, 87, 125]) may be the most popular way for practitioners to interact with technical explanations. This approach seeks to communicate how small changes in specific features may lead to changes in specific model outputs.

Local Interpretable Model-agnostic Explanations (LIME). LIME [125], one of the leading algorithms using the local feature importance paradigm, aims to "explain the pre-

[16] Consider our prior analogy of an automotive system. Just as a diagnostic code on a modern automobile can be used to determine whether a car's check engine light is on because of an electrical problem in the sensor, or because of a legitimate engine malfunction, a tradition within computer science identifies an explanation with an outcome that enables other trained technical experts to debug, assess, or otherwise improve a faulty system. Nevertheless, interpreting a car's diagnostic code, and knowing what to do next, usually requires some specialized expertise (either mechanical, computational, or both – or sometimes, just deep familiarity with the system).

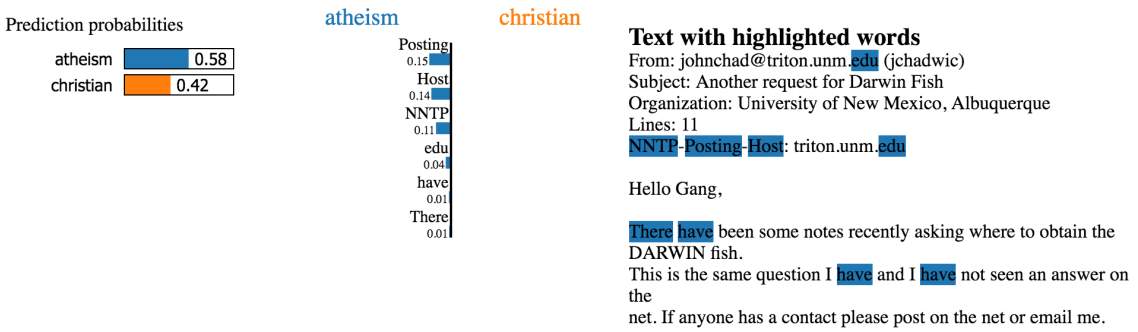


Fig. 5. An example of output from LIME which emphasizes text features that led a specific paragraph to be classified as about atheism, rather than Christianity. The original image may be found at this URL: https://github.com/marcotcr/lime/blob/master/doc/images/twoclass.png

dictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model...by presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction." LIME can help developers to understand how changes in individual features might change the model's output around a specific prediction. To the extent that these insights generalize, and are based on meaningful features, they may help developers to infer the model's causal mechanisms; however, these approaches may also mislead if they become subject to spurious correlations. For example, Figure 5 shows the output of LIME when applied to a paragraph of text that was classified as about atheism rather than about Christianity. This classifier appears to focus on properties of the author (e.g., the fact that they originate from an academic institution, as indicated by the .edu in their email address) and specific stylometric features (e.g., use of the words "have" and "there") rather than words that might be indicative of content.

In so doing, LIME draws users' attention to specific features that the model uses to make a specific prediction, thus connecting a specific output to a simplified representation of the model that generated that output. For example, Figure 6 demonstrates how a classifier designed to tell the difference between wolves and huskies classified a particular image based on the presence of snow in the background (and not based on the anatomical features that would actually differentiate these two species). A data scientist with appropriate domain knowledge would be able to use this information to modify or otherwise debug this faulty classification.

Thus, this process bears some resemblance to the definition of explanation presented above; however, there are also important differences. First, LIME does not provide the user with an explanation of the model *per se*, but rather provides the users with a simplified model that approximates the more complex model that the algorithm is trying to explain. In effect, LIME replaces a complex, causal, description of a model's inner workings with a simpler description of a different model whose results are only correlated with the original model. For example, LIME provides no information regarding whether the wolf vs. husky

(a) Husky classified as wolf (b) Explanation

Fig. 6. An example of output from LIME which emphasizes input features that are diagnostic of a specific image classification. In this case, a picture of a husky was incorrectly classified as a wolf because of the presence of snow in the background. Such information may help tool developers debug overfit classifiers. This image was originally presented in [125]

classifier in Figure 6 would make accurate predictions on images that did not have snow in their backgrounds.

The authors of LIME argue that these simplified models (e.g., regression models with a small number of coefficients) are inherently more interpretable because they "provide qualitative understanding between the input variables and the response." While this goal is broadly consistent with Fuzzy-Trace Theory's definition of gist, gists, when educated, capture a human expert's insight regarding which features are most likely to generalize. Techniques such as LIME may aid humans in generating these representations, and indeed, preliminary experiments seem to suggest that human subjects could use these techniques to remove features that interfered with predictive accuracy – i.e., they could make a better classifier – and that a small sample of human subjects with with data science expertise (and, in particular, familiarity with the concept of spurious correlation) might be able to use LIME to derive better explanations.

SHapley Additive exPlanations (SHAP). Like LIME, the SHAP [87] family of models start from the premise that "The best explanation of a simple model is the model itself" and attempts to therefore represent complex models with simpler models. SHAP therefore returns importance scores for each feature, which are analogous to regression coefficients. For a given prediction, SHAP scores indicate how much any of these features contributed to that prediction, as in Figure 7.

As such, SHAP models possess many of the same strengths and weaknesses of LIME,

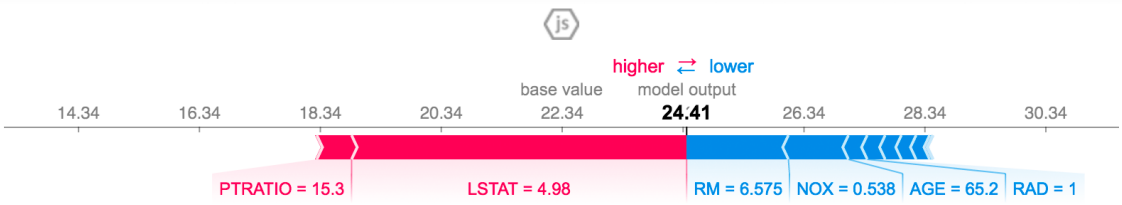


Fig. 7. An example of output from SHAP which indicates which indicates the model's baseline value, the marginal contributions of each of its features, and the final prediction. Such an approach is analogous to a graphical interpretation of linear regression coefficients. The original image may be found at https://github.com/slundberg/shap

albeit with the ability to generalize to a larger class of machine-learning models. These models are verbatim in the most concrete sense – they output a set of rules (feature importance scores), which may be applied in a rote manner to generate a post-hoc description of the desired prediction. However, they do not communicate causal mechanisms, and they are prone to unknown error as the model is applied outside of the local neighborhood of a specific prediction. Individual human subjects – such as informed practitioners – that have the willingness and the ability to examine these findings in depth may be able to leverage their own background knowledge to generate an explanation, but SHAP does not provide enough information to help these practitioners figure out when the model no longer applies. In effect, these techniques provide human users with a stimulus that they must then explain or interpret whereas true "black box" models do not even provide this stimulus.

Explainable Neural Networks. Whereas SHAP and LIME seek to explain complex models using a regression-like paradigm (i.e., a linear additive function), Explainable Neural Networks (XNNs) [144] use a more general formulation based on an "additive index model" [127]. Here, the algorithm seeks to return a function that describes how model predictions vary with changes to individual parameters (or, more recently, pairs of parameters [148]). As in LIME and SHAP, these models can help data scientists with the appropriate training to understand how changing a specific feature might change the model's prediction, albeit at the risk of inferring spurious correlations. These approaches have especially been applied to deep neural network models, in which one neural network is used to provide a simplified representation of another, and then rendered into a table that is analogous to an Analysis of Variance, showing main effects and, in some cases, two-way interactions [35].

However, LIME is not without limitations: the explanations that analysts might draw from applying these tools may, themselves, be based on spurious correlations or may engender false confidence in model predictions outside the scope of the immediate neighborhood of the datapoint that LIME is attempting to explain[17]. Worse, these misleading explanations may be engineered by adversaries seeking to take advantage of humans' tendency to

[17]Precisely this sort of inappropriate model generalization led NASA engineers to conclude that an impact from foam insulation upon the space shuttle *Columbia*'s wing from its external tank did not pose a threat, precipitating the loss of the space shuttle upon re-entry into Earth's atmosphere [16]

impute causality where none exists [134].

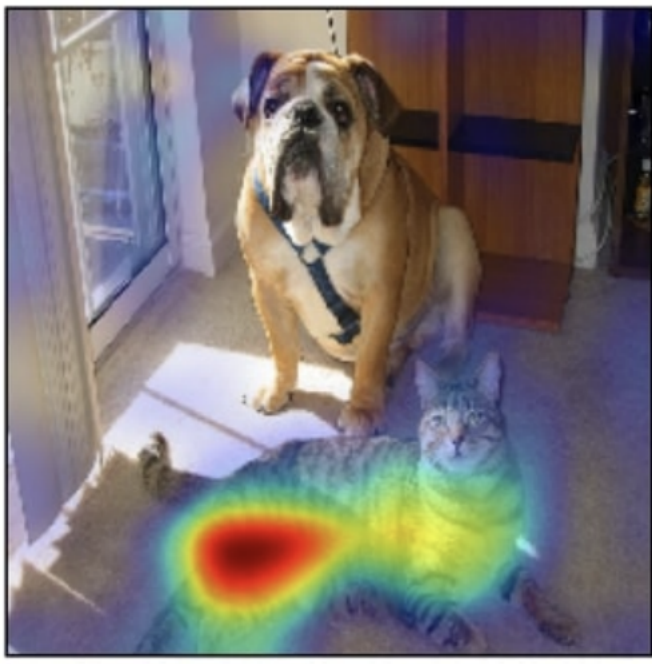
Gradient-weighted Class Activation Mapping (Grad-CAM). Grad-CAM is a method designed to explain computer vision models that use deep learning architectures (specifically, convolutional neural nets – currently the state-of-the-art architecture for computer vision). Specifically, Grad-CAM "uses the gradients of any target concept (say 'dog' in a classification network or a sequence of words in captioning network) flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept" [129]. Grad-CAM takes advantage of the layered architecture of CNNs to identify those regions of an image that are most diagnostic of a particular prediction. For example, Figure 8 shows how Grad-CAM output can draw a user's attention to the part of an image that are diagnostic of a specific prediction that a user wishes to explain. This is a visual version of the feature importance paradigm – where the features are ensembles of specific pixels – with several of the corresponding strengths and limitations.

3.2.2 "Simpler Models Are Inherently More Interpretable"

Rudin [128] has sharply critiqued techniques which seek to generate simple explanations of complex models, arguing that they can obfuscate the actual inner-workings of these models in a manner that misleads decision-makers and analysts. Models that are locally-accurate do not provide information on the extent of that accuracy or whether its degradation is graceful or sudden. Instead of trying to approximate more complex models with simpler models, Rudin argues that simpler models should be used directly, because they are more "interpretable" (i.e., by data scientists), especially when the stakes are high. The rationale for this approach is that data scientists, at least, may understand the model's inner workings.

Scalable Bayesian Rule Lists. Scalabale Bayesian Rule Lists [147] are one example of a technique that aims to avoid model complexity. In contrast to the techniques outlined above, which seek to provide continuous representations of complex models, Scalable Bayesian Rule Lists explicitly do not attempt to compete with "black box classifiers such as neural networks, support vector machines, gradient boosting or random forests. It is useful when machine learning tools are used as a decision aid to humans, who need to understand the model in order to trust it and make data-driven decisions." As such, SBRLs do not aim to achieve both high predictive accuracy and explainability; rather, they seek to provide a set of simplified (verbatim) probabilistic rules that can be used to partition a datset (see Table 2).

Generalized Additive Models with pairwise interactions. One approach that may address Rudin's critique relies on using Generalized Additive Models with pairwise interactions (GA^2Ms) – a class of models which restrict the "contribution of a single feature to the final prediction" to depend only on that feature [34]. The intent of these models is to disentangle each feature from all other features such that they may be evaluated independently of one another. Figure 9 shows the output of an GA^2M applied to a dataset that predicts risk of hospital 30-day admission for pneumonia. As above, these models are



(c) Grad-CAM ‘Cat’

(i) Grad-CAM ‘Dog’

Fig. 8. An example of output from Grad-CAM, indicating which pixels in an image are diagnostic of the predicted class (dog or cat). The original image may be found at [129]

Table 2. Example of SBRL output, which seeks to explain whether a customer will leave the service provider. PP = Probability that the label is positive. Source: [147]

Rule-list	PP	Test Accuracy
if (Contract=One year&StreamingMovies=Yes),	0.20	0.81
else if (Contract=Two year),	0.032	0.98
else if (Contract=One year),	0.054	0.97
else if (tenure<1year&InternetService=Fiber optic),	0.70	0.72
else if (PaymentMethod=Electronic check & InternetService=Fiber optic),	0.48	0.45
else (TechSupport=No&OnlineSecurity=No),	0.42	0.64
else (default),	0.22	0.78

primarily correlational in nature and may help domain experts to select features – for example, the authors note that pneumonia readmission risk decreases, rather than increases, with asthma – a counterintuitive finding. This model surfaces that finding. However, domain experts must then explain that finding post-hoc, as follows:

> [P]atients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit). The good news is that the aggressive care received by asthmatic pneumonia patients was so effective that it lowered their risk of dying from pneumonia compared to the general population. The bad news is that because the prognosis for these patients is better than average, models trained on the data incorrectly learn that asthma lowers risk, when in fact asthmatics have much higher risk (if not hospitalized)[34].

The discussion above indicates that these concerns apply to explainability – where the goal is to assist a data scientist to understand how a model works – but may apply less to interpretability, where the aim is fundamentally to assist a decision maker to link a model output to a meaningful distinction that allows them to leverage their values, goals, and preferences to make a choice. Specifically, the explanation given above might help a user to debug the model, or even to decide whether or not to trust the model; however, it may not explicitly provide the user with meaningful information that can inform their ultimate treatment decision.

Monotonically Constrained Gradient-Boosting Machines Gradient-Boosting Machines seek to use an ensemble of "weak learners" – i.e., models with low predictive accuracy – to jointly make accurate predictions. This approach leads to significant improvements in predictive capability, at the cost of model complexity. In order to cope with this complexity, Monotonically Constrained Gradient-Boosting Machines impose a constraint that any given feature in the model must have a monotonic relationship with the output. This is theorized to increase explainability because these monotonic relationships restrict the relationship between features and predictions to have clear qualitative directions – an increase in the feature must consistently lead to either an increase or decrease in the

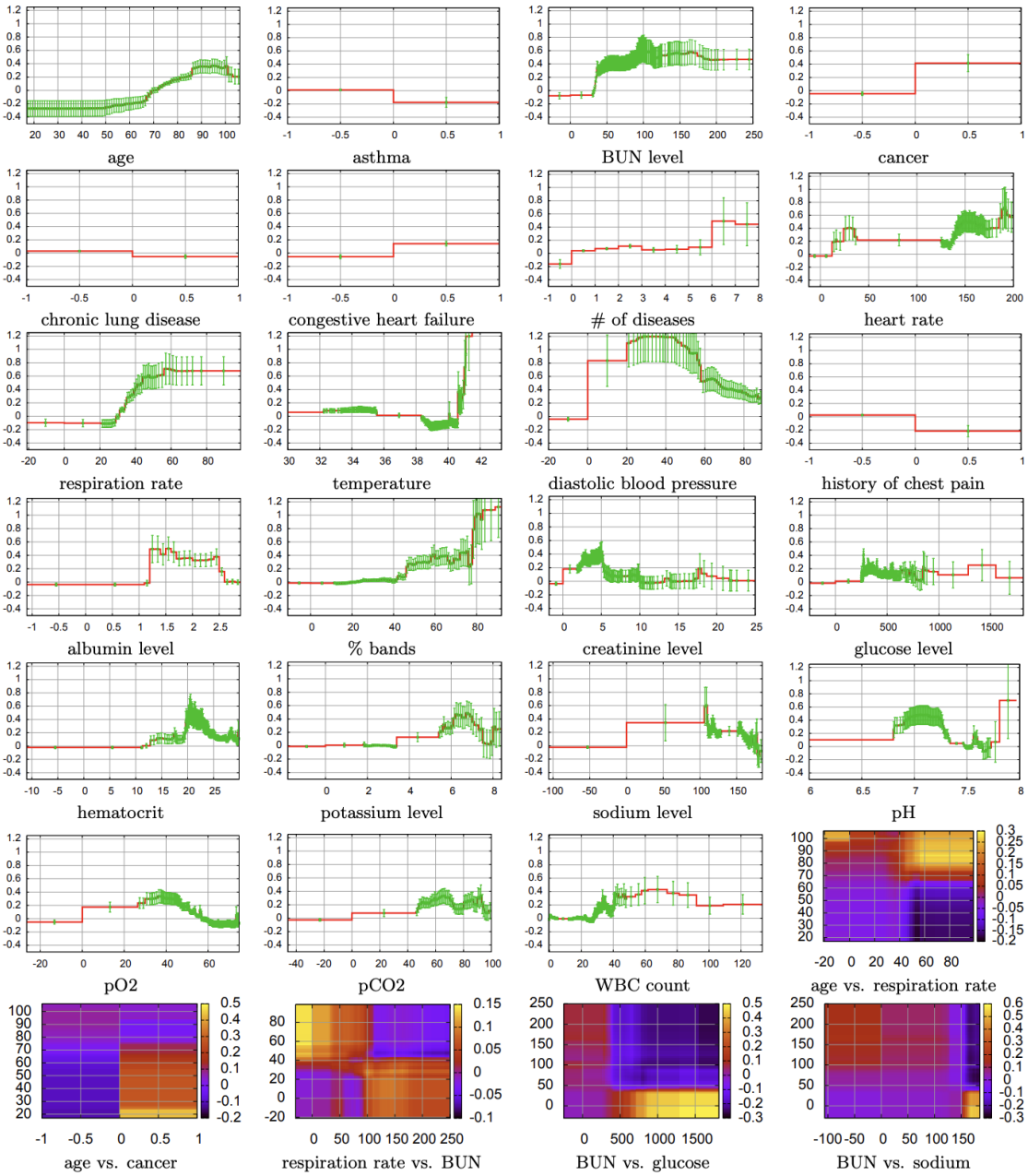


Fig. 9. An example of output from a GA^{2}M which indicates how several features (horizontal axes) vary with relative risk of readmission for pneumonia at 30 days (vertical axis). Pairwise interactions are shown in the heatmaps at the bottom of the figure. The original image is in [34].

prediction. As above, these models assume that simpler functional forms are inherently more explainable. However, these models, in their current form, may simply apply a form of regularization that is not necessarily grounded in domain knowledge. Monotonicity may be appropriate in some cases – such as a dose-response curve – but not in others – such as in modeling waves or other sinusoidal behavior. Domain knowledge is required to determine whether monotonicity constraints, or any other constraints, are appropriate. Absent that domain knowledge, application of such constraints may indeed simplify the model, but may do so in a misleading way that can foster inference of incorrect explanations.

3.2.3 Limitations of Current Explainable AI Models

Generally speaking, the assumption that simplified models are inherently interpretable assumes some degree of domain knowledge on the part of model users – i.e., that they have sufficient data science expertise to make sense of linear models, decision trees, rule lists, etc. Furthermore, these "interpretable" models may not provide users with sufficient context to apply their values, goals, and principles to enable a decision. These techniques are truly verbatim in the sense that they provide a rule but with no insight as to the algorithm's actual mechanism. They provide correlation but not causation. However, they may assist subject-matter experts or data scientists to infer causation. These techniques can encourage experts with appropriate background expertise to more deeply explore the mechanisms by which a particular classification was made, although without making those mechanisms themselves explicit. Thus, a technical expert can perhaps leverage their background knowledge of the type of algorithm used to infer causality from these tools. This may enable them to construct an explanation in the same way that a juror or a reader can infer coherent structure from cohesive text. However, it is ultimately the human that imputes the explanation to the model output. The techniques outlined above do not provide explicit representations of causal mechanisms or interface with users' values, goals, or preferences. Rather, they must rely on humans' background knowledge for their utility. Thus, these models assume much of the viewer, including potentially-significant domain knowledge regarding the meaning of technical terms (such as "hematocrit" in the GA^2M pneumonia diagnosis example), the ability to distinguish between continuous and discrete variables, etc. Similarly, subjects must have extensive subject-matter expertise to be able to recognize, for example, that prior history of asthma should not be associated with lower pneumonia risk. Thus, the model, on its own, is not interpretable or explainable in the way that psychologists conceive these terms, but may help users with the appropriate background knowledge, and willingness to investigate, to draw more meaningful and accurate conclusions.

Because these models are correlational in nature, they may be subject to spurious association. Indeed, it has long been acknowledged in the social sciences [130] that the identification of meaningful structure in data (e.g., due to a correlation or regression), is only the first step in the imputation of a causal mechanism and, absent a counterfactual (such as an experimental control group), cannot be relied upon to make causal claims. Thus, approaches that simplify complex models by reducing them to a set of monotonic

relationships may mislead users into imputing a causal mechanism within the model where none exists. This problem is not limited to computational systems, but is a general feature of complex engineered system with multiple interacting parts [29]. Thus, future work in explainable artificial intelligence may productively focus on how to help data scientists and domain experts to accurately impute causal claims while avoiding drawing inferences based on spurious correlation.

3.2.4 Purpose-Built Graphical User Interfaces

In general, the approaches outlined above seek to enhance explainability by helping users to understand how changes to a specific feature might change model output. Although these are theorized to enhance explainability when a data scientist can use them to infer causal mechanisms about how the algorithm works, these techniques may be less effective for establishing interpretability – i.e., meaning in context for the end user. Whereas developers need to know *how* the system works so that they can identify flaws in their implementation and fix them, members of the public or subject-matter experts from other fields typically lack the in-depth technical training and expertise of computer scientists; nor should they be expected to develop it. For example, an immigration lawyer may want to know the legal implications of a visa review algorithm or a financial analyst may want to know the financial implications of a credit rating algorithm. Often, these users will simply assume that the algorithm was implemented correctly, and that the training data were adequately representative. Finally, job/visa/credit applicants would naturally want to know the standards by which they are being assessed and whether they are competitive for a specific position. These users need to know *why* a system generated its result. That is, they seek to *make sense* of model output such that they can contextualize it in terms that are *meaningful* to them.

In some cases, Graphical User Interfaces (GUIs) such as Google's What-If tool, may be paired with model output to assist users with limited numeracy or statistical knowledge to "get the gist". For example, there is a body of work in medical decision-making examining individual differences in graph literacy (e.g., [48]) and techniques that may be used to overcome these differences to communicate the gist of complex medical information [37, 37, 145]. However, designers must take care not to assume that a graphical format is necessarily more interpretable. Rather, the graphical output must be contextualized with appropriate representations of base rates, thresholds, and other indicators of meaningful categorical distinctions that, in many cases, may need to be elicited from users. Ultimately, machine-generated interpretations must be contextualized in terms of background knowledge and goals, and tailored to individual differences, if they are to be effective.

3.2.5 Coherent Topic Models

Topic models are a family of Bayesian inference algorithms that have been widely applied to text data for information retrieval and document summarization [13]. The most widely-applied of these algorithms, Latent Dirichlet Allocation (LDA) [14], infers latent "topics"

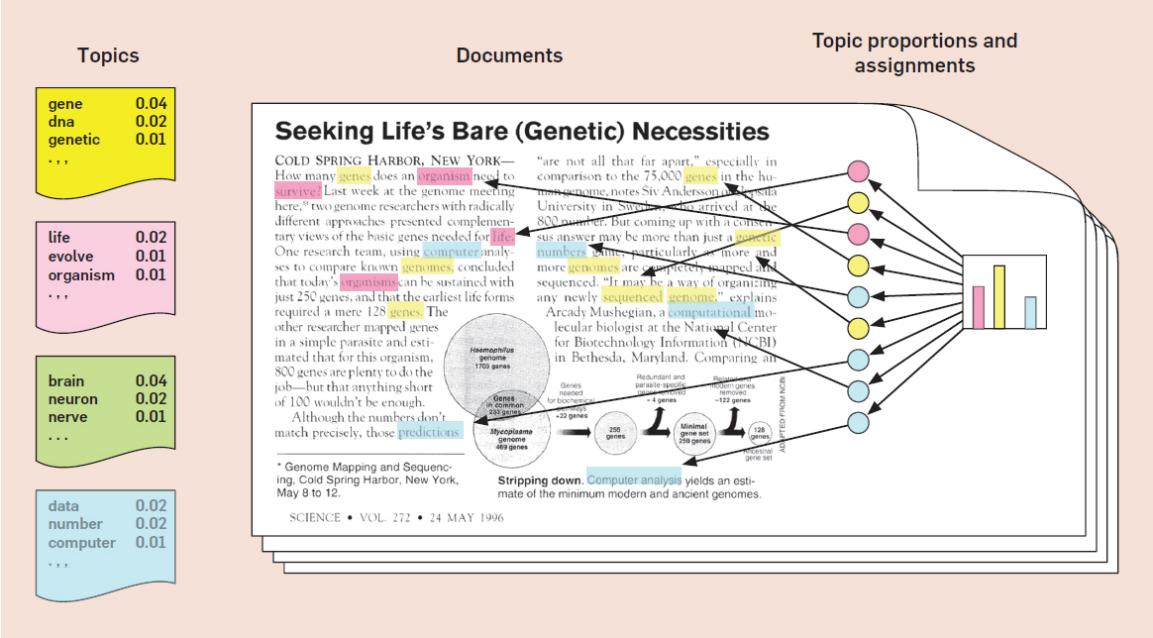


Fig. 10. A visualization of Latent Dirichlet Allocation output from [13]. Probabilistic topic models such as LDA map each word in a text corpus to a topic. The most frequent words in that topic are then presented to humans for interpretation.

that are supposed to contain semantic content held in common across multiple documents. In practice, these topics are actually a probability distribution over words in the text corpus upon which the LDA model is trained. Humans use topic models by inspecting the top words, or top documents, for any given topic and then assigning meaning to those topics [59], with some even going so far as to claim that topic models explicitly measure the gist of text [60]. However, more recent work has shown that humans have difficulty interpreting some topic model output [36], especially when they are not familiar with how the algorithm works [83]. Although computer scientists have developed measures to improve the coherence (hypothesized to increase interpretability) of topic model output [81, 96], the resulting output does not explicitly provide an interpretation to human users, but remains a list of words with associated topic probabilities, which humans must interpret (see Figure 10). Nevertheless, topic models are perhaps unique among ML algorithms in that their users have attempted to explicitly engineer interpretability into their structure and output using tasks that are evaluated by non-expert humans with no knowledge of how the algorithm works. Future work should focus on evaluating this approach and potentially applying it to other algorithmic paradigms.

4. Incorporating Insights from Psychology Into Design

How might one evaluate the explainability and interpretability of AI systems in a manner that is psychologically plausible? How might we design systems that satisfy these psycho-

logical definitions? This section addresses these questions.

4.1 Psychological Correlates of AI Expert System Paradigms

The ML techniques described in section 3.2 reflect a tension between two different approaches to evaluating the quality of mathematical models, and "rational" behavior more broadly.

4.1.1 Coherence and "White-Box Models"

The first approach, which Hammond [62] called "coherence", emphasizes the *process* by which a result is obtained. According to this approach, a result is judged according to whether it is obtained by following logical rules that start from universally-accepted axioms. Early AI systems – especially rule-based systems – exhibited high degrees of coherence and, by extension, explainability per the psychology-based definitions in this paper. Strengths of the coherence approach include its guarantees of logical completeness – if the axioms are correct, and the rules are followed unerringly, then the conclusions must necessarily be correct. However, these systems have been criticized for their fragility in real-world decision-making (e.g., the work that inspired the GA^2M approach [33]). In practice, they may fail if the axioms are not correct (but at least one could determine how that conclusion was attained!) For example, a classical expert system is typically constructed by eliciting rules from experts; however, these rules could be applied "mindlessly" (e.g., without relevant background knowledge, such as about time, human anatomy, or important exceptions, as in the case of a rule-based medical expert system [93]). To the extent that those rules are correct, the system's recommendations should be correct; however, the process of eliciting these rules may introduce sources of error that would invalidate the results, such as when patients do not disclose all relevant information to an algorithm since they do not know the algorithm requires it, or because they do not trust the algorithm to use that information appropriately. Indeed, traditional rule-based AI systems are marked by a strict adherence to verbatim rules that occasionally lead to the wrong conclusions. Attempts to *oversimplify* machine learning models based on purely algorithmic considerations may actually impose harmful biases in some circumstances [77].

White-box models. Like human decision processes emphasizing coherence, "white-box" ML models are transparent with humans able to readily understand how they operate because they follow a set of transparent rules. Examples of white-box models include linear models, which can be readily transformed from input to prediction by multiplying by well-defined coefficient values. These models also seem to accord with Rudin's [128] definition of interpretability. Furthermore the explainable AI techniques outlined in section 3.2 appear to be designed to make black-box models more like white-box models (at the risk of introducing potential spurious correlations).

4.1.2 Correspondence and "Black-Box Models"

It is generally assumed that explainability and predictive accuracy must be traded against one another. Consistent with this perceived dichotomy, Hammond [62] defined "correspondence" approaches as those which emphasizes empirical accuracy all else. Here, a decision is considered to be good if it leads to a good result, regardless of how this result is obtained. This is analogous to the machine learning paradigm, which emphasizes prediction over explanation [131, 149]. Standard machine learning techniques aim to optimize specific predictive metrics, such as accuracy, precision, recall, F-score, etc. Furthermore, any number of algorithms may be employed regardless of whether the underlying theory of the algorithm is a good description of the process generating the data. This approach is consistent with Hammond's definition of correspondence since it privileges predictive accuracy over a specific causal theory. Deep neural nets, in particular, have been criticized – but also lionized – because they often achieve significant predictive predictive performance at the cost of explainability. Thus, like ML, the weaknesses of the correspondence approach are fundamentally connected to low explainability – a method may achieve the right answers for the wrong reasons – i.e., dut to spurious correlation – thus, there is no confidence that future model results will be correct. As Hammond [62] states,

> Scientific research seeks both coherence and correspondence, but gets both only in advanced, successful work. Most scientific disciplines are forced to tolerate contradictory facts and competitive theories...But policymakers find it much harder than researchers to live with this tension because they are expected to act on the basis of information. (p. 54).

Black-box models. Like decision approaches favoring correspondence, "black-box models" are those whose inner workings are inaccessible, and hence incomprehensible, to human users becuase they emphasize predictve accuracy over explainability. These models can only be evaluated for their predictive qualities, and one must simply "trust" that they will continue to perform in the real world as they do on training data. Prototypical examples of black-box models include deep neural nets.

Hammond's discussion highlights that the current tension between explanation and prediction in machine learning and statistics (see also [131]) is, in fact, a longstanding feature of the scientific method that can nevertheless be at odds with policy and legal requirements for data-driven decision-making. Indeed, there seems to be a common perception that models possessing high correspondence are likely to have low coherence and vice versa (although see [97]). However, the discussion above highlights that explanations are fundamentally about providing coherent outputs that describe the *process* by which a model achieved a given result. In contrast, interpretations emphasize how a stimulus (either a model's output, a datapoint or dataset, or a description of the model itself) is contextualized in the broader world setting, and thus could be evaluated relative to correspondence criteria.

4.1.3 A Third Way: Enhancing Interpretability by Querying Human Experts and "Grey-Box Models"

By distinguishing between interpretation and explanation, we propose that human experts' gists may be thought of as analogous to a "grey box model" – one for which a full mechanistic explanation (i.e., a white box model) is not available, but for which blind trust (i.e., a black box model) is also not required. This middle road is achieved by experts' communicating the *gists* of their decision-making processes, rather than trying to explain all of the details of their structured mental models. Specifically, experts can communicate how what they are doing is consistent with the values of users in easy-to-understand categorical terms while not necessarily possessing the ability to describe the exact mechanisms in every detail. We propose grey box model design as a goal for interpretable AI.

Human decision-making exhibits varying degrees of coherence and correspondence. Specifically, experts' gist representations tend to correspond to better real-world outcomes, exhibiting correspondence; however, experts may violate coherence [7] – i.e., they can provide an explanation for an action in a particular context, but that explanation may not necessarily generalize to all contexts, roughly analogous to a linear estimate of a nonlinear model. Unlike these linear estimates which provide explanatory power over a narrow parameter range, experts can be queried for their rationales. The above discussion emphasizes how, far from being a unique feature of machine learning models, high correspondence with low explainability may also be a feature of some types of human expertise. In fact, there is a significant body of literature in engineering management that discusses the "tacit" nature of human expertise [100–102, 105, 106]. In other words, like the most complex models, human experts may not be consciously aware of *how* they have obtained a certain outcome. Nevertheless, they are often able to describe *why* they did what they did – for example expert tennis players were more likely justify their actions relative to the goals of the game whereas novices focused more conscious attention on the mechanics of executing specific maneuvers [92]. Thus, experts' decisions show a high degree of empirical correspondence despite being subject to "biases" under predictable circumstances [112]. Furthermore, subject matter experts tend to rely on, consume, and prefer to use model output that *interprets*, rather than explains, relevant results.

This discussion suggests that designers may be able to enhance interpretability by creating "grey-box models", that can provide the rationale for a given decision relative to a set of functional requirements. In section 4.2, we argue that this goal is aligned with has long dealt with similar problems when attempting to integrate large-scale systems from across several different complex domains of expertise into a common artifact to be used by consumers, including policymakers, with varying levels of technical sophistication [21].

4.2 Interpretable and Explainable Outputs Are Different Abstractions of a System

Arguably, designing an explainable or interpretable AI system boils down to selecting the appropriate level of abstraction at which to communicate system output given a user's needs. Here, we review research in both psychology and systems engineering research

supporting this claim. This literature primarily focuses on drawing an analogy between the concept of mental representation and levels of abstraction.

Computer science depends on successive levels of abstraction for its successful operation simply because the operations of computational systems are far too complex, at the physical level, to explain to even the most expert computer scientists. Consider that a computer is itself an abstraction – a "Turing machine" – that is implemented using another abstraction – bits – that are themselves implemented in silicon semiconductors. Abbott [4, 5] points out that bits themselves are abstractions and, to the extent that the abstraction isn't violated (e.g., because of physical limits such as too much heat), system developers do not need to understand the mechanism (e.g., the physical regularities) underlying the implementation of the computational system that they are using. Similar logic applies to software development. Although it is certainly helpful, in some cases, to understand computer architecture when designing software, most software developers do not need a detailed understanding of the code implementation underlying a particular computer's operating system when designing an application. Kroll [79] even argues that explanations that focus on the *mechanisms* of an AI system's operations actually *obscure* users' abilities to understand how the system operates within its social context (e.g., power structure). Indeed, users routinely utilize applications – e.g., on the Internet – without detailed explanations of how those applications work. Rather, they are familiar with a set of functions that the application is intended to perform and, as well as those functions are carried out in a way that does not impose undue externalities, the user generally does not need, or even care to, know about implementation details. It is precisely these externalities that lie at the heart of the need for system interpretability. Thus, Abbott provides a section pertaining to "platform governance" that provides an overview of some of the research on governance of common resources that could productively be adapted to algorithmic explainability and interpretability, and especially to the development of standards in this area. Specifically, these standards might be framed in terms of high-level *requirements* [40] with measures of effectiveness for interpretable and explainable systems.

4.2.1 Psychological Evidence that Abstraction Improves Interpretability and Decision Quality

The distinction between "how" and "why" levels of mental representation is theoretically-motivated and empirically-supported in the psychology literature. Construal Level Theory (CLT) [52, 139], a leading theory of abstraction in human psychology, distinguishes between mental representations in terms of their "psychological distance" (although see [120] for an alternative account). Here, a less distant representation entails memory for things that are either spatially or temporally proximate, whereas a more distant representation entails memory for things that are more distant. According to CLT, questions that ask "how" – what we have called explanations – are more psychologically proximate than are questions that ask "why" – interpretations. Furthermore, CLT characterizes the distinction between more and less distant representations in terms of their level of abstraction [52].

Finally, recent findings [47] show that increasing psychological distance, including by providing more abstract representations – lead to better decisions because these more distant representations make use of gist interpretations. This has direct implications for design: interpretable AI systems – i.e., those that help end users to make meaningful decisions – may, in fact, be *less* explainable, and vice versa, at least for end users without data science expertise.

4.2.2 Abstraction Hierarchies in Engineering

The idea that explanations – justified relative to system implementation – and interpretations – justified relative to system goals – are distinct is also supported by extensive literature in human factors engineering. This paper has described how Fuzzy-Trace Theory posits that these representations are encoded distinctly and in parallel, with experts preferring to rely on more abstract descriptions. This section discusses how these notions of abstraction are used in engineering design by practicing engineers. Specifically, CLT's core construct – the abstraction hierarchy distinguishing between more detailed "how" questions – that are defined relative to specific implementations and thus less meaningful with respect to system goals – and less detailed, but more meaningful "why" questions – was presaged in the human factors engineering/systems engineering literature by the work of Jens Rasmussen [108] (see Figure 11). Rasmussen's work, which was conducted exclusively on technical experts (largely electrical engineers making consequential decisions pertaining to complex systems, such as nuclear power plants), defined an abstraction hierarchy as one in which "...the functional properties of a technical system are represented in several levels of functional abstraction along the means-end dimension" [108](p. 235) – a continuum of representations bounded below by the system's "physical form" (here, analogous to the actual physical implementation of a machine-learning algorithm on a computer, in terms of which bits are flipped), through to a system's "generalized" and "abstract function" (analogous to the software functions implemented by a specific ML system, which give outputs justified relative to those functions), which define the system's causal structure and/or information flow topology, and bounded above by the system's "functional purpose" – i.e., its objectives relative to its end users. Rasmussen and Lind [109] further note that, when coping with complexity, engineers often rely on multiple levels of representation, even switching between them for the purpose of diagnoses, with the higher levels indicating the functional requirements for why the lower levels are implemented, and the lower levels indicating the specific concrete realization of how the higher level requirements are carried out. Thus, users and designers can only understand system output to the extent that they possess the technical or domain expertise required to utilize corresponding levels of the abstraction hierarchy. In general, moving between levels of abstraction requires understanding that level of abstraction on its own terms and bringing to bear relevant background knowledge.Thus, the form of abstraction that we discuss here bears some similarity to Kolmogorov's [78] definition of complexity – it is a simplified representation that can be expressed using a short description, where the description length is defined relative to a pre-defined knowl-

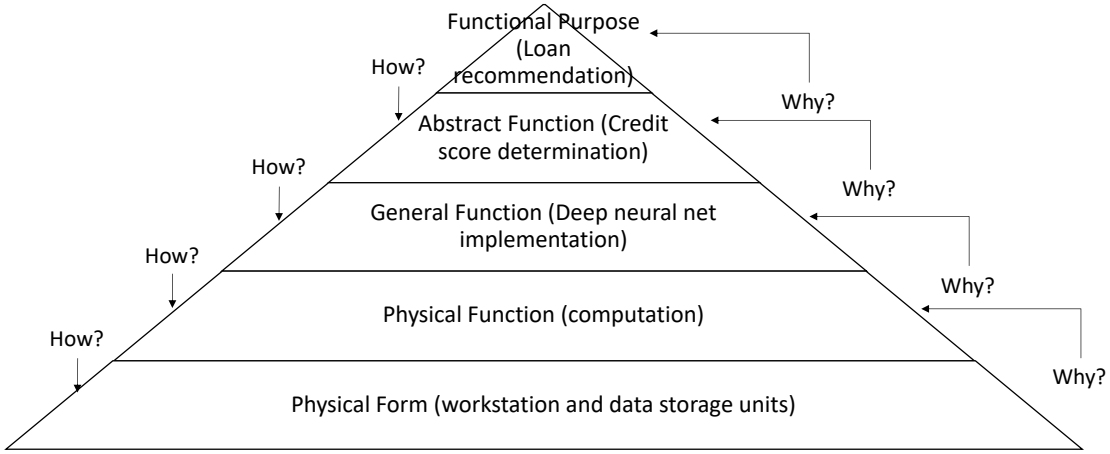


Fig. 11. An example of an Abstraction Hierarchy for a ML system designed to make loan recommendations. Each higher level implemented by the level immediately below it and each lower level implements a technical solution to carry out a function specified by the higher level.

edge base (usually, a programming language, but see also [20, 45, 91]). However, the term *abstraction* is itself polysemous (for a review, see [30]) and we emphasize that the form of abstraction that we refer to is not just simplification, but rather is augmented by the user's background knowledge [120].

4.2.3 Design for Interpretability and Explainability as Requirements Engineering

CLT and the Abstraction Hierarchy both suggest that systems should be evaluated in terms of their "requirements", rather than their specific technical specifications (see also [79]. Good requirements are "solution neutral" – they do not specify details of implementation, but only the function that a system should carry out. The aim is to specify this function as clearly as possible while not overly constricting the technical design of the system itself. For example, a solution neutral requirement might specify that a system should achieve an accuracy of at least 0.85 without specifying the particular algorithm used. Thus, requirements are specified in terms of measures of effectiveness, or measures of performance,

that have a meaning in terms of the system's ultimate function; not its implementation. This form of design is familiar to machine learning researchers who frequently use several different algorithms to obtain the same function. For example, the function may be classification, which has well-defined metrics – precision, recall, accuracy, etc. – with the best metric selected based on the task. Given the function required, and the metric benchmark that the system is supposed to achieve (i.e., the requirement), system designers may use any number of algorithms to achieve this function. For classification, candidate algorithms might include logistic regression classifiers, Naive Bayes classifiers, support vector classifiers, k-nearest neighbor classifiers, convolutional neural nets, etc., all perform the same function, yet using very different implementations.

The concept of solution neutrality is equally familiar to legal scholars and regulators and legal scholars, who have significant experience with evaluating complex systems, such as human-drug interactions, in other highly technical domains [80, 140]. Thus, the key to a good design is matching the system's evaluation metric to the end-user's goals. Standard ML metrics, such as those defined above, may be insufficient to address the system's functional requirements. Systems engineering, and especially requirements engineering, is the discipline that focuses on evaluating end-user needs and translating these into a suite of metrics. However, for sufficiently complex systems, groups of engineers must pool their knowledge to achieve better design outcomes. Attempts to generate unified models seamlessly integrating input from different engineering fields ("Model-Based Systems Engineering") have yielded mixed results [21]. In contrast, techniques that seek to translate meaningful information between engineering specialties are both more widely accepted by practicing engineers. Although there is no guarantee that these techniques generate optimal outcomes, they are, at a minimum, acceptable [22, 69, 70]. Thus, other fields of engineering have developed methods to ensure that expert input from multiple fields can be integrated into a larger, complex project. Specifically, systems engineering is the field of engineering that is concerned with coordinating these multiple experts from multiple domains. Since, nowadays, no one person can be an expert in all fields of human inquiry, systems engineers have developed a set of tools and techniques to increase the confidence with which expertise is deployed when developing complex systems.

5. Conclusion

Our review indicates that explainability and interpretability should be distinct requirements for ML systems, with the former primarily being of value to developers who seek to debug systems or otherwise improve their design, while the latter more useful to regulators, policymakers, and general users. However, this dichotomy is not absolute – individual differences may be associated with reliance on one representation of a system more so than another. For developers, the explanation may be very similar to the interpretation – similar to how numerate individuals derive decisions that are informed by verbatim calculations – however, if developers lack domain expertise, they may be unable to contextualize their interpretations in terms that are meaningful to end users. Similarly, end users lacking de-

velopers' expertise may be unable to understand how the system arrived at its conclusion (but they also may not desire to).

5.1 Implications for Designing Explainable and Interpretable Artificial Intelligence

The history of modern engineering demonstrates that systems can be designed to facilitate both explainability and interpretability and, indeed, there are examples of such throughout recent history, ranging from drug approvals to the automotive sector. Although significant effort has been devoted to developing automated approaches to create explanations of AI algorithms, comparatively little attention has focused on interpretability. Since interpretations differ between individuals, more research is needed in order to determine how best to link model output to specific gists so that users can appropriately contextualize this output. The extent to which this process can be fully automated, or would require curation by domain experts, remains an open question.

The above discussion indicates that *interpretable algorithms are those that contextualize data by placing it in the context of structured background knowledge, representing it in a simple manner that captures essential, insightful distinctions, and then justifying the corresponding output relative to values elicited from human users*. Such representations contextualize the model's output and provide meaning to the human user in terms of values stored in long-term memory. Typically, these values (or similar preference-generating constructs, such as goals) cannot be elicited directly from data based on rote, brittle, verbatim association. Thus, techniques to simplify complex models are likely to share this brittleness [98]. In contrast, gist representations are simple, yet flexible and insightful; they bring to bear contextual elements – such as goals and values – that are not explicitly represented in the data. Future work may therefore productively focus on eliciting these gist representations from experts in the form of mappings from structured background knowledge to meaningful, yet simple, categories – associated with goals, values, principles, or other preferences. Previous "expert systems" lacked the ability to scale precisely because of the difficulty eliciting these essential distinctions [33]. To move beyond this impasse, the discussion in this paper highlights the need to account for *multiple levels of mental representation* when generating interpretable AI output. In short, rather than assimilating human cognition to machine learning, we might benefit from designing machine learning models to better reflect empirical insights about human cognition [116]. In between a verbatim, data-driven approach, and an inflexible top-down schematic approach lies one in which human users engage in a process of contextualizing model output that is then used to select between, and refine, existing background knowledge structures. There is some preliminary evidence that these "communicative" approaches — in which human users interact with and curate the output of AI systems — may show some promise. Furthermore, users' needs vary with individual differences, e.g., in metacognition and training. Future work should therefore focus on characterizing these factors within user communities. This review provides the theoretical basis for such an approach, and provides explicit directions for future work: such approaches must communicate the *gist* of the data to the user.

Acknowledgments

The authors thank Andrew Burt, Carina Hahn, Patrick Hall, Sharon Laskowski, P. Jonathon Phillips, Mark Przybocki, Valerie F. Reyna, Reva Schwartz, Brian , and Paul Witherell for their insightful comments and discussions.

References

[Com]

[Cen] Centor score (modified/mcisaac) for strep pharyngitis - mdcalc. https://www.mdcalc.com/centor-score-modified-mcisaac-strep-pharyngitis. (Accessed on 10/07/2020).

[Wor] Worried your sore throat may be strep? — cdc. https://www.cdc.gov/groupastrep/diseases-public/strep-throat.html. (Accessed on 10/07/2020).

[4] Abbott, R. (2006). Emergence explained: Abstractions: Getting epiphenomena to do real work. *Complexity*, 12(1):13–26.

[5] Abbott, R. (2007). Putting complex systems to work. *Complexity*, 13(2):30–49.

[6] Adadi, A. and Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6:52138–52160.

[7] Adam, M. B. and Reyna, V. F. (2005). Coherence and correspondence criteria for rationality: Experts' estimation of risks of sexually transmitted infections. *Journal of Behavioral Decision Making*, 18(3):169–186.

[8] Alba, J. W. and Hasher, L. (1983). Is memory schematic? *Psychological Bulletin*, 93(2):203.

[Ammermann] Ammermann, S. Adverse action notice requirements under the ecoa and the fcra - consumer compliance outlook: Second quarter 2013 - philadelphia fed.

[10] Axelrod, R. (2015). *Structure of decision: The cognitive maps of political elites*. Princeton university press.

[11] Bhatt, U., Xiang, A., Sharma, S., Weller, A., Taly, A., Jia, Y., Ghosh, J., Puri, R., Moura, J. M. F., and Eckersley, P. (2019). Explainable machine learning in deployment. *arXiv:1909.06342 [cs, stat]*. arXiv: 1909.06342.

[12] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018). "it's reducing a human being to a percentage"; perceptions of justice in algorithmic decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, page 1–14. arXiv: 1801.10408.

[13] Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

[14] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[15] Bless, H., Betsch, T., and Franzen, A. (1998). Framing the framing effect: The impact of context cues on solutions to the 'asian disease'problem. *European Journal of Social Psychology*, 28(2):287–291.

[16] Board, U. S. C. A. I. (2003). *Columbia Accident Investigation Board Report*. Columbia Accident Investigation Board. Google-Books-ID: J7F7c4kRy_wC.

[17] Brainerd, C. J. and Reyna, V. F. (2001). Fuzzy-trace theory: Dual processes in memory, reasoning, and cognitive neuroscience.

[18] Brainerd, C. J. and Reyna, V. F. (2002). Fuzzy-trace theory and false memory. *Current Directions in Psychological Science*, 11(5):164–169.

[19] Brainerd, C. J. and Reyna, V. F. (2005). *The science of false memory*, volume 38. Oxford University Press.

[20] Briscoe, E. and Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, 118(1):2–16.

[21] Broniatowski, D. A. (2018a). Building the tower without climbing it: Progress in engineering systems. *Systems Engineering*, 21(3):259–281.

[22] Broniatowski, D. A. (2018b). Do design decisions depend on "dictators"? *Research in Engineering Design*, 29(1):67–85.

[23] Broniatowski, D. A., Klein, E. Y., May, L., Martinez, E. M., Ware, C., and Reyna, V. F. (2018). Patients' and clinicians' perceptions of antibiotic prescribing for upper respiratory infections in the acute care setting:. *Medical Decision Making*.

[24] Broniatowski, D. A., Klein, E. Y., and Reyna, V. F. (2014). Germs are germs, and why not take a risk? patients' expectations for prescribing antibiotics in an inner-city emergency department. *Medical Decision Making*.

[25] Broniatowski, D. A. and Reyna, V. F. (2013a). Gist and verbatim in narrative memory. In *2013 Workshop on Computational Models of Narrative*, volume 32, page 43–51.

[26] Broniatowski, D. A. and Reyna, V. F. (2013b). Gist and verbatim in narrative memory. In *2013 Workshop on Computational Models of Narrative*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

[27] Broniatowski, D. A. and Reyna, V. F. (2018). A formal model of fuzzy-trace theory: Variations on framing effects and the allais paradox. *Decision*, 5(4):205.

[28] Broniatowski, D. A. and Tucker, C. (2017a). Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Systems Engineering*, 20(6):483–496.

[29] Broniatowski, D. A. and Tucker, C. (2017b). Assessing causal claims about complex engineered systems with quantitative data: internal, external, and construct validity. *Systems Engineering*, 20(6):483–496.

[30] Burgoon, E. M., Henderson, M. D., and Markman, A. B. (2013). There are many ways to see the forest for the trees: A tour guide for abstraction. *Perspectives on Psychological Science*, 8(5):501–520.

[31] Cacioppo, J. T., Petty, R. E., Feinstein, J. A., and Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119(2):197–253.

[32] Cacioppo, J. T., Petty, R. E., and Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, 48(3):306–307.

[33] Caruana, R. (2017). Intelligible Machine Learning for Critical Applications Such As

Health Care. aaas.

[34] 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*, page 1721–1730.

[35] Chen, J., Vaughan, J., Nair, V. N., and Sudjianto, A. (2020). Adaptive explainable neural networks (axnns). *arXiv:2004.02353 [cs, stat]*. arXiv: 2004.02353.

[36] Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., and Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 1–12, Glasgow, Scotland Uk. Association for Computing Machinery.

[37] Cozmuta, R., Wilhelms, E., Cornell, D., Nolte, J., Reyna, V., and Fraenkel, L. (2018). Influence of explanatory images on risk perceptions and treatment preference. *Arthritis care & research*, 70(11):1707–1711.

[38] Curseu, P. L. (2006). Need for cognition and rationality in decision-making. *Studia Psychologica*, 48(2):141.

[39] De Weck, O. L., Roos, D., and Magee, C. L. (2011). *Engineering systems: Meeting human needs in a complex technological world*. Mit Press.

[40] Dick, J., Hull, E., and Jackson, K. (2017). *Requirements engineering*. Springer.

[41] Diehl, J. J., Bennetto, L., and Young, E. C. (2006). Story recall and narrative coherence of high-functioning children with autism spectrum disorders. *Journal of abnormal child psychology*, 34(1):83–98.

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

[43] Edwards, L. and Veale, M. (2018). Enslaving the algorithm: From a "right to an explanation" to a "right to better decisions"? *IEEE Security Privacy*, 16(3):46–54.

[44] Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., and Smith, D. M. (2007). Measuring numeracy without a math test: development of the subjective numeracy scale. *Medical Decision Making*, 27(5):672–680.

[45] Feldman, J. (2009). Bayes and the simplicity principle in perception. *Psychological Review*, 116(4):875.

[46] Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives*, 19(4):25–42.

[47] Fukukura, J., Ferguson, M. J., and Fujita, K. (2013). Psychological distance can improve decision making under information overload via gist memory. *Journal of Experimental Psychology: General*, 142(3):658.

[48] Galesic, M. and Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, 31(3):444–457.

[49] Gallo, D. (2013). *Associative illusions of memory: False memory research in DRM and related tasks*. Psychology Press.

[50] Gernsbacher, M. A. (1996). *The structure-building framework: What it is, what it*

might also be, and why, page 289–311. Psychology Press.

[51] Gernsbacher, M. A., Varner, K. R., and Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3):430.

[52] Gilead, M., Trope, Y., and Liberman, N. (2020). Above and beyond the concrete: The diverse representational substrates of the predictive brain. *Behavioral and Brain Sciences*, 43:e121.

[53] Gilovich, T., Griffin, D., and Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press.

[54] 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. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

[55] Gleaves, L. P., Schwartz, R., and Broniatowski, D. A. (2020). The role of individual user differences in interpretable and explainable machine learning systems. *arXiv preprint arXiv:2009.06675*.

[56] Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., and Holzinger, A. (2018). Explainable ai: The new 42? In Holzinger, A., Kieseberg, P., Tjoa, A. M., and Weippl, E., editors, *Machine Learning and Knowledge Extraction*, Lecture Notes in Computer Science, page 295–303. Springer International Publishing.

[57] Goldman, S. R., McCarthy, K. S., and Burkett, C. (2015). 17 interpretive inferences in literature. *Inferences during reading*, page 386.

[58] Graesser, A. C., Singer, M., and Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological review*, 101(3):371.

[59] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235.

[60] Griffiths, T. L., Steyvers, M., and Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological review*, 114(2):211.

[61] Hall, P., Gill, N., and Schmidt, N. (2019). Proposed guidelines for the responsible use of explainable machine learning. *arXiv:1906.03533 [cs, stat]*. arXiv: 1906.03533.

[62] Hammond, K. R. (2000). Coherence and correspondence theories in judgment and decision making. *Judgment and decision making: An interdisciplinary reader*, page 53–65.

[63] Hans, V. P., Helm, R. K., and Reyna, V. F. (2018). From meaning to money: Translating injury into dollars. *Law and human behavior*, 42(2):95.

[64] Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and Clancey, W. J. (2018). Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems*, 33(3):87–95. Publisher: IEEE.

[65] Hoffman, R. R. and Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, 32(3):68–73. Publisher: IEEE.

[66] Hoffman, R. R., Mueller, S. T., and Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4):78–86. Publisher: IEEE.

[67] Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable ai: Challenges and prospects. *arXiv:1812.04608 [cs]*. arXiv: 1812.04608.

[68] Jones, N. A., Ross, H., Lynam, T., Perez, P., and Leitch, A. (2011). Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16(1).

[69] Katsikopoulos, K. V. (2009). Coherence and correspondence in engineering design: informing the conversation and connecting with judgment and decision-making research. *Judgment and Decision Making*, 4(2):147–153.

[70] Katsikopoulos, K. V. (2012). Decision methods for design: insights from psychology. *Journal of Mechanical Design*, 134(8):084504–1–084504–4.

[71] Kintsch, W. (1974). The representation of meaning in memory.

[72] Klein, E. Y., Martinez, E. M., May, L., Saheed, M., Reyna, V., and Broniatowski, D. A. (2017). Categorical risk perception drives variability in antibiotic prescribing in the emergency department: A mixed methods observational study. *Journal of General Internal Medicine*.

[73] Klein, G. (2008). Naturalistic decision making. *Human factors*, 50(3):456–460.

[74] Klein, G. (2018). Explaining explanation, part 3: The causal landscape. *IEEE Intelligent Systems*, 33(2):83–88. Publisher: IEEE.

[KLEIN et al.] KLEIN, G., HOFFMAN, R., and MUELLER, S. Naturalistic Psychological Model of Explanatory Reasoning: How People Explain Things to Others and to Themselves.

[76] Klein, G. A., Orasanu, J., Calderwood, R., Zsambok, C. E., et al. (1993). *Decision making in action: Models and methods*. Ablex Norwood, NJ.

[77] Kleinberg, J. and Mullainathan, S. (2019). Simplicity creates inequity: implications for fairness, stereotypes, and interpretability. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 807–808.

[78] Kolmogorov, A. N. (1965). Three approaches to the definition of the concept "quantity of information". *Problemy peredachi informatsii*, 1(1):3–11.

[79] Kroll, J. A. (2018). The fallacy of inscrutability. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376(2133):20180084.

[80] Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., and Yu, H. (2016). Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3):633–706.

[81] Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

[82] LeBoeuf, R. A. and Shafir, E. (2003). Deep thoughts and shallow frames: On the susceptibility to framing effects. *Journal of Behavioral Decision Making*, 16(2):77–92.

[83] Lee, T. Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., and Findlater, L. (2017). The human touch: How non-expert users perceive, interpret, and fix topic models. *International Journal of Human-Computer Studies*, 105:28–42.

[84] Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., and Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of behavioral decision making*, 25(4):361–381.

[85] Linderholm, T., Everson, M. G., van den Broek, P., Mischinski, M., Crittenden, A., and Samuels, J. (2000). Effects of causal text revisions on more- and less-skilled readers' comprehension of easy and difficult texts. *Cognition and Instruction*, 18(4):525–556.

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

[87] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

[88] Malle, B. F. (2010). Intentional action in folk psychology. *A companion to the philosophy of action*, pages 357–365.

[89] Mandler, J. M. (1983). What a story is. *Behavioral and Brain sciences*, 6(04):603–604.

[90] Marti, D. and Broniatowski, D. A. (2020). Does gist drive nasa experts' design decisions? *Systems Engineering*.

[91] Mathy, F. and Feldman, J. (2012). What's magic about magic numbers? chunking and data compression in short-term memory. *Cognition*, 122(3):346–362.

[92] McPherson, S. L. and Thomas, J. R. (1989). Relation of knowledge and performance in boys' tennis: Age and expertise. *Journal of experimental child psychology*, 48(2):190–211.

[93] Miller, R. A., Pople Jr, H. E., and Myers, J. D. (1982). Internist-i, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine*, 307(8):468–476.

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

[95] Miller, T., Howe, P., and Sonenberg, L. (2017). Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *arXiv:1712.00547 [cs]*. arXiv: 1712.00547.

[96] Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272.

[97] Mittelstadt, B., Russell, C., and Wachter, S. (2019a). Explaining explanations in ai. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 279–288. Association for Computing Machinery.

[98] Mittelstadt, B., Russell, C., and Wachter, S. (2019b). Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pages 279–288, Atlanta, GA, USA. Association for Computing Machinery.

[99] Moray, N. (1990). A lattice theory approach to the structure of mental models. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 327(1241):577–583.

[100] Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Orga-*

nization science, 5(1):14–37.

[101] Nonaka, I. and Takeuchi, H. (1995). *The knowledge-creating company: How Japanese companies create the dynamics of innovation*. Oxford university press.

[102] Nonaka, I., Toyama, R., and Konno, N. (2000). Seci, ba and leadership: a unified model of dynamic knowledge creation. *Long range planning*, 33(1):5–34. 21.

[103] Pennington, N. and Hastie, R. (1993). *The story model for juror decision making*. Cambridge University Press Cambridge.

[104] Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., and Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17(5):407–413.

[105] Polanyi, M. (1962). Tacit knowing: Its bearing on some problems of philosophy. *Reviews of modern physics*, 34(4):601. 20.

[106] Polanyi, M. (1967). The tacit dimension.

[107] 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.

[108] Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision-making and system management. *IEEE Transactions on systems, man, and cybernetics*, (2):234–243.

[109] Rasmussen, J. and Lind, M. (1981). Coping with complexity. Technical Report Risø-M-2293, Risø National Laboratory, Roskilde, Denmark.

[110] Reese, E., Haden, C. A., Baker-Ward, L., Bauer, P., Fivush, R., and Ornstein, P. A. (2011). Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of Cognition and Development*, 12(4):424–462.

[111] Regulation, G. D. P. (2018). General data protection regulation (gdpr). *Intersoft Consulting, Accessed in October 24*, 1.

[112] Reyna, V. (2018). When irrational biases are smart: A fuzzy-trace theory of complex decision making. *Journal of Intelligence*, 6(2):29.

[113] Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current directions in psychological science*, 13(2):60–66.

[114] Reyna, V. F. (2008). A theory of medical decision making and health: fuzzy trace theory. *Medical decision making*, 28(6):850–865.

[115] Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision making*.

[116] Reyna, V. F. (2020). Of viruses, vaccines, and variability: Qualitative meaning matters. *Trends in Cognitive Sciences*.

[117] Reyna, V. F. and Adam, M. B. (2003). Fuzzy-trace theory, risk communication, and product labeling in sexually transmitted diseases. *Risk Analysis*, 23(2):325–342. 29.

[118] Reyna, V. F. and Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and individual Differences*, 7(1):1–75.

[119] Reyna, V. F. and Brainerd, C. J. (1998). Fuzzy-trace theory and false memory: New frontiers. *Journal of experimental child psychology*, 71(2):194–209.

[120] Reyna, V. F. and Broniatowski, D. A. (2020). Abstraction: An alternative neurocog-

nitive account of recognition, prediction, and decision making. *Behavioral and Brain Sciences*, 43.

[121] Reyna, V. F., Chick, C. F., Corbin, J. C., and Hsia, A. N. (2014). Developmental reversals in risky decision making intelligence agents show larger decision biases than college students. *Psychological Science*, 25(1):76–84.

[122] Reyna, V. F. and Farley, F. (2006). Risk and rationality in adolescent decision making implications for theory, practice, and public policy. *Psychological science in the public interest*, 7(1):1–44.

[123] Reyna, V. F. and Lloyd, F. J. (2006). Physician decision making and cardiac risk: effects of knowledge, risk perception, risk tolerance, and fuzzy processing. *Journal of Experimental Psychology: Applied*, 12(3):179.

[124] Reyna, V. F., Lloyd, F. J., and Brainerd, C. J. (2003). Memory, development, and rationality: An integrative theory of judgment and decision making. *Emerging perspectives on judgment and decision research*, pages 201–245.

[125] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[126] Ribeiro, M. T., Singh, S., and Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *AAAI*, volume 18, pages 1527–1535.

[127] Ruan, L. and Yuan, M. (2010). Dimension reduction and parameter estimation for additive index models. *Statistics and its Interface*, 3(4):493–499.

[128] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *arXiv:1811.10154 [cs, stat]*. arXiv: 1811.10154.

[129] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2019). Grad-cam: Visual explanations from deep networks via gradient-based localization. *arXiv:1610.02391 [cs]*. arXiv: 1610.02391.

[130] Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Wadsworth Cengage learning.

[131] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310. Zbl: 1329.62045.

[132] Simonite, T. (2020). How an algorithm blocked kidney transplants to black patients. *Wired*.

[133] Singer, M. and Remillard, G. (2008). Veridical and false memory for text: A multiprocess analysis. *Journal of Memory and Language*, 59(1):18–35.

[134] Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. *arXiv:1911.02508 [cs, stat]*. arXiv: 1911.02508.

[135] Slovic, P., Finucane, M. L., Peters, E., and MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis: An International Journal*, 24(2):311–322.

[136] Tomsett, R., Braines, D., Harborne, D., Preece, A., and Chakraborty, S. (2018).

Interpretable to whom? a role-based model for analyzing interpretable machine learning systems. *arXiv:1806.07552 [cs]*. arXiv: 1806.07552.

[137] Trabasso, T., Secco, T., and Van Den Broek, P. (1984). *Causal cohesion and story coherence.*, page 83–110. Lawrence Erlbaum Associates.

[138] Trabasso, T. and Van Den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of memory and language*, 24(5):612–630.

[139] Trope, Y. and Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological review*, 117(2):440.

[140] Tutt, A. (2017). An fda for algorithms. *Admin. L. Rev.*, 69:83.

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

[142] Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323.

[143] Van den Broek, P. (2010). Using texts in science education: Cognitive processes and knowledge representation. *Science*, 328(5977):453–456.

[144] Vaughan, J., Sudjianto, A., Brahimi, E., Chen, J., and Nair, V. N. (2018). Explainable neural networks based on additive index models. *arXiv preprint arXiv:1806.01933*.

[145] Wilhelms, E. A., Reyna, V. F., Brust-Renck, P., Weldon, R. B., and Corbin, J. C. (2015). Gist representations and communication of risks about hiv-aids: A fuzzy-trace theory approach. *Current HIV Research*, 13(5):399–407.

[146] Wolfe, C. R., Widmer, C. L., Reyna, V. F., Hu, X., Cedillos, E. M., Fisher, C. R., Brust-Renck, P. G., Williams, T. C., Vannucchi, I. D., and Weil, A. M. (2013). The development and analysis of tutorial dialogues in autotutor lite. *Behavior research methods*, 45(3):623–636.

[147] Yang, H., Rudin, C., and Seltzer, M. (2017). Scalable bayesian rule lists. In *International Conference on Machine Learning*, pages 3921–3930. PMLR.

[148] Yang, Z., Zhang, A., and Sudjianto, A. (2020). Enhancing explainability of neural networks through architecture constraints. *IEEE Transactions on Neural Networks and Learning Systems*.

[149] Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6):1100–1122.

[150] Zsambok, C. E. and Klein, G. (2014). *Naturalistic decision making*. Psychology Press.

[151] Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological bulletin*, 123(2):162.