DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY

Explaining by Conversing: The Argument for Conversational XAI Systems

Citation

Marrakchi, Wassim. 2021. Explaining by Conversing: The Argument for Conversational XAI Systems. Bachelor's thesis, Harvard College.

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37368579

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Explaining by Conversing

The Argument for Conversational XAI Systems

A DISSERTATION PRESENTED
BY
WASSIM MARRAKCHI
TO
THE DEPARTMENT OF COMPUTER SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF ARTS
IN THE SUBJECT OF
COMPUTER SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MARCH 2021

©2014 – Wassim Marrakchi
all rights reserved.

Thesis advisor: Professor Elena Glassman Wassim Marrakchi

Explaining by Conversing
The Argument for Conversational XAI Systems

Abstract

The interest in chatbots and conversational agents is as old as artificial intelligence (AI) itself. [171,176] Recently, multiple members of the HCI community including *Weld and Bansal* (2018) [177] have suggested that conversational explanation systems is the best path forward for explainable human-agent interaction. This recommendation is often presented without its supporting arguments so we embarked on this thesis to shed some light on the call behind conversational explainable AI (XAI) systems. First, we survey the research on the need for explanations from AI systems and on the models' ability to provide them. Second, we provide a set of obstacles in the way of interpreting and making meaning of these explanations and explain these obstacles by drawing from the results of several studies in human-computer interaction, machine learning, cognitive science, and education theory. Finally, we take these obstacles into account to argue for conversational explanation systems and propose a Wizard-of-Oz (WoZ) experiment to test some of our hypotheses.

Contents

Listing of figures

To my parents Maledh and Samia, my brother Mehdi, and my family away from home: Andrew, Evan, and Joey.

Acknowledgments

I would like to thank Professors Cynthia Dwork, Eddie Kohler, and Mike Smith for their academic advising over the years, as well as Professor Elena Glassman for her patience, support, and meticulous feedback in this thesis process and in the various research projects that motivated this work. I would also like to express my gratitude to the Hazem and Karen Ben-Gacem Undergraduate Scholarship Fund for supporting my financial aid award in the past four years and making my time at Harvard possible.

0

Introduction

As opaque AI systems are increasingly being employed in critical contexts, the demand for understanding their inner workings is increasing from various stakeholders and is fueling more and more research on explainable AI (XAI) systems.[11,139] In the absence of a clear taxonomy, the terms "interpretability" and "explainability" are often used synonymously in the existing literature.[17,79,91,103] However, while these two notions are related, they actually mean different things when talking about AI systems. Explainability is about the extent to which explainability methods capture the internal in-

ner workings of a model. As defined by *Chakraborty et al.* (2017), the explainability of a model refers to "the type and completeness of the output given when a model is queried for reasoning behind its decision."[25] On the other hand, interpretability focuses on the meanings explanations lead to and describes the extent to which the explanations of these internal inner-workings are comprehensible to a user agent.

Given these differences, the assessment of explainability and interpretability relies on different interdependent metrics,[25] and one might think of interpretability as a bigger umbrella than explainability. *Fürnkranz et al.* (2019) distinguish between three aspects of interpretability: syntactic interpretability, epistemic interpretability, and pragmatic interpretability.[44] Syntactic interpretability encompasses what *Bibal and Frenay* (2016) refer to as mental fit and is concerned with the user's ability to comprehend the knowledge encoded in the model and the provided explanations.[16,44] Epistemic interpretability encompasses notions such as trustworthiness and credibility and corresponds to an assessment of the extent to which a model and its explanations are in line with existing domain knowledge and correlated with the user's prior knowledge.[44] Pragmatic interpretability ecompasses notions such as interestingness, usability, and acceptability and captures the extent to which a model and its explanations can serve their intended purpose and satisfy the end-user's needs.[44] Note that these aspects are not independent; syntactic interpretability can be a prerequisite to the other two notions,[16,44] and epistemic interpretability can be a prerequisite to pragmatic interpretability.[44]

In this thesis, we focus on the notion of interpretability and the meanings explanations lead to. We provide some background on explainability and its motivations but choose to focus on how these explanations are understood by human beings. Working off of the obstacles in the way of interpretability, we join the members of the HCI community who have called for the development of conversational explanation systems and provide a number of arguments for what conversational XAI systems may be able to do right. We conclude this work with a proposal for a Wizard-of-Oz user study that will test some of these hypotheses. This proposal is the fruit of an on-going collaboration with Elizabeth

Hu under the supervision of Elena Glassman. We also present preliminary results from previous work with Nari Johnson and Elizabeth Hu that motivates and supports this proposal.

Our Contributions

In Chapter 1, we cover both the motivations for providing explainability and the technical considerations for extracting explanations from models. In Section 1.1, we provide a short exposition of the history of the field of artificial intelligence following *Russell and Norvig*'s own exposition [153] and focus on the turning points that led to the current obfuscation in mainstream models. In Section 1.2, we survey the latest research and real-world applications to establish the need for explainability and expose the reach of these algorithms and their faults. We also summarize the argument of *Doshi-Velez et al.* (2019) for society's need for explanations in decision-making contexts. [35] In Section 1.3, we provide an exposition of the technical considerations for extracting explanations from both transparent models and black-box models.

In Chapter 2, we turn to the obstacles in the way of making meaning out of these explanations. In Section 2.1, we provide an exposition of *Doshi-Velez and Kim*'s taxonomy for the different approaches to evaluating a model's comprehensibility and situate the focus of this thesis within this taxonomy. [33] In Section 2.2, we argue that monolithic interpretability is a myth because a variety of factors influence it including the context and the audience. In Section 2.3, we draw on research from cognitive science and education theory to argue that cognitive biases play an important role in a model's interpretability. In Section 2.4, we discuss the side effects of explanation tools on a model's interpretability. In Section 2.5, we will argue for a new understanding of a human-computer interaction with an explainable AI (XAI) system and suggest viewing it as a bidirectional communication.

In Chapter 3, we argue for conversational explanation systems and propose an experiment to assess participants' satisfaction with such systems, elicit their implicit expectations, and investigate its

ability to mitigate cognitive errors. In Section 3.1, we present the design recommendation of *Weld and Bansal* (2018) that inspired our work and its motivations.[177] In Section 3.2, we defend their call for conversational explainable AI (XAI) systems by laying out some hypotheses for how these systems can hold up against the obstacles outlined in Chapter 2. In Section 3.3, we describe a proposal for a Wizard-of-Oz study to test the effectiveness of conversational explanation systems and present some preliminary results supporting this effectiveness. The study proposal is based on our on-going collaboration with Elizabeth Hu and the preliminary results are from our previous work with Elizabeth Hu and Nari Johnson. Both research projects were supervised by Elena Glassman.

"In many ways, the modern theory of computation is the long-awaited science of the relations between parts and wholes; that is, of the ways in which local properties of things and processes interact to create global structures and behaviors."

Marvin Minsky

1

The Why and the How of Explainability

In this chapter, we will focus on the notion of explainability, its motivations, and its technical considerations. In Section 1.1, we provide a short overview of the history of artificial intelligence (AI) and focus on the turning points that led to the obfuscation of AI models. In Section 1.2, we show the reach of artificial intelligence in our lives, provide evidence for their faultiness, and summarize the instances in which explanations may be expected from AI systems as outlined by *Doshi-Velez et al.* (2019).[35] In Section 1.3, we turn to the technical considerations behind extracting explanations from

transparent models and black-box models and give some background on the state-of-the-art explainability methods.

1.1 A Short History of Artificial Intelligence

A plethora of ideas from a variety of fields inspired the early research into thinking machines. Aristotle's informal system of syllogisms for proper reasoning,[99] George Boole's propositional logic,[18] Charles Babbage's Analytical Engine,[19] Ada Lovelace's ideas for Babbage's machine,[99] and Norbert Wiener's cybernetics[151] are but a few of the influential contributions that laid the groundwork for modern artificial intelligence.[153] Although the early work of McCulloch and Pitts from 1943 on a model of artificial neurons was the first description of what is now known as neural networks,[99] it is Alan Turing who is considered the father of artificial intelligence for his avant-garde vision. As early as 1947, Turing held lectures on thinking machines at the London Mathematical Society and his work culminated in the 1950 landmark article "Computing Machinery and Intelligence" where he introduced the Turing Test, machine learning, reinforcement learning, and many other influential ideas.[153,171]

In 1956, the term "Artificial Intelligence" was coined by James McCarthy at the Dartmouth conference, the "official birthdate of the new science" and the field of Artificial Intelligence (AI) research.[30,112,153] The workshop's attendees included Allen Newell, Herbert Simon, Marvin Minsky, and Arthur Samuel.[153] Along with their students, these scientists ended up dominating the field for the 20 subsequent years.[153] The early enthusiasm generated by this workshop led to a period full of successes with a promising performance of early rule-based AI systems on simple examples.[153] These early approaches, now called weak methods, attempted to build general-purpose search mechanisms that could find complete solutions from elementary reasoning steps.[153] However, it rapidly became clear that simple syntactic manipulations, brute-force techniques, and basic structures like percep-

trons were doomed for failure with larger and more difficult problems given the daunting combinatorial explosion and the limited computational power.[4,153]

By going back to the drawing board, the next generation of Artificial Intelligence (AI) systems focused on more powerful, domain-specific knowledge and created the methodology of expert systems for more narrow areas of expertise.[153] These expert systems are able to solve complex problems by reasoning through bodies of knowledge in the form of facts and rules.[75] By their very nature, these systems are inherently interpretable and don't present any explainability barriers.[11] The first knowledge-based systems were created in the 1970s and included the Dendral program for inferring molecular structure from the information provided by a mass spectrometer,[21] the MYCIN program for diagnosing blood infections,[153] and the SHRDLU system for understanding natural language.[181] The approach's many successes led to an increase in the demand for workable knowledge representation schemes and the deployment of the first commercial expert system at the Digital Equipment Corporation in 1982.[113,153] This frenzy is considered the first "hype" of AI.[11,153]

From 1986 onwards, the second "hype" of AI came with the return of the neural networks of McCulloch and Pitts and the back-propagation learning algorithm of Bryson and Ho made possible by the increase in computational power.[152,153] During these days, the field of artificial intelligence also abandoned its isolationism from the rest of computer science.[153] As researchers adopted the scientific method to empirically confirm hypotheses,[28] they started to better understand the problems' complexities and paved the way for today's research agendas.[153] However, with the increasing availability of large data sources, they started paying more attention to the data than to the algorithmic solutions themselves as soon as they realized that mediocre algorithms with large training data can outperform the best known algorithms with smaller datasets.[9,60,63,153] This realization shifted the focus of many solutions to more obscure learning methods instead of hand-coded knowledge engineering.[60]

1.2 Why Do We Need Explainability?

In 1957, Herbert Simon, a pioneer of artificial intelligence, predicted that, in a "visible future," the variety of problems machines can handle will be "coextensive with the range to which the human mind has been applied." [160] While Simon's "visible future" took a bit longer than he expected, 50 years later, this future is materializing with the growing ubiquitousness of AI. Today, artificial intelligence is taking over driving, [169] speech recognition, planning and scheduling, [83] game championships, [53] logistics planning, [32] and many other areas of daily life. All over the world, algorithms are distributing our vaccines, [61] managing our pandemics, [56,120,121] waging our wars, [1] tracking our employees, [43,64,104] enforcing our laws, [6,90,123,148] supporting our clinical decisions, [47] and even feeding our gambling tendencies. [74] Many of these AI models perform really well based on their performance but focusing on a single metric (e.g., classification accuracy) is "an incomplete description of most real-world tasks." [34]

1.2.1 The Faults in our Models

By relying on artificial intelligence (AI) to fulfill many of these functions, we've given obscure learning methods more and more control over our lives and we're only just starting to see beyond the tip of the iceberg. Some of society's most critical AI applications have turned out to be prone to concerning failures and mistakes, going against some of our most important values and principles. The 2016 study of *Angwin et al.* on the criminal risk assessment tool COMPAS sounded a wake-up call for the AI community when they found that its predictions were unreliable and racially biased. [7] The 2018 study of *Gebru and Buolamwini* demonstrated the frightening race and gender bias in commercially available facial recognition software. [22] A 2021 study by *Juneja and Itra* show problematic patterns of vaccine-misinformation amplification on our e-commerce platforms. [82] A full survey of these failures is beyond the scope of this thesis but the dangers of unchecked AI applications are undeniable.

Many of these problems are inherent to our algorithms. Deep neural networks (DNNs) have

been easily fooled into misclassifying inputs with no resemblance to the true category.[127] One-pixel-attacks and other techniques are able to change a network's classification of any image to any target class by making imperceptible alterations to a small number of pixels.[119,133,166] State-of-the-art character-level and word-level DNN-based text classifiers and natural language networks can also be manipulated.[80,100] To this day, researchers are still discovering new ways in which our favorite algorithms can fail to meet our expectations,[23,52,67,107] and it's become clear that the faults in our models are hard to unearth because of their obfuscated inner-workings.

1.2.2 When Should We Expect an Explanation?

To be clear, it's unrealistic to expect an explanation from every AI system. In fact, even before the advent of automated decision-making systems, human decision-makers have not had to provide an explanation for every single decision they made. Explanations can reduce the time and effort available to spend on other tasks, be used in a socially irresponsible way to game the system in the presence of a mismatch between the goals of the parties involved (e.g., credit scoring), or decrease observers' trust in some decisions.[35,116] However, there are instances in which the benefits of an explanation outweigh its costs and where a decision-maker is morally, socially, or legally obligated to provide it.

Doshi-Velez et al. (2019) outline three factors that society does take into consideration when requiring explanations of the decision-making.[35] The first factor is the impact of the decision on persons other than the decision-maker.[35] The second factor is the possibility of acting on the explanation and correcting for an error in past or future decision-making.[35] The third factor is a belief that an error has occurred in the decision-making process informed by knowledge of inadequate inputs, inexplicable outcomes, or nonalignment of the decision-maker's interests with society's.[35] While these factors may seem exhaustive, it is important to note that, depending on the situation, they may be present in varying degrees and may not account for the decision-maker's own interests (e.g., increase trust in the decision-making process).[35]

When providing legal accountability, the legal system can require a party to "provide an explanation for a decision when the opposing party has provided some degree of proof that the decision caused a legal-cognizable and redressable injury."[35] This requirement can cover administrative agencies,[87] private decision-makers in certain industries,[114] or even individual litigants on a case by case basis.[163] The study of *Doshi-Velez et al.* (2019) shows that the requirement to explain under the law is present in several countries with small variations in the explanations' role, who is obligated to provide them, and the amount of evidence needed to bring about the requirement.[35]

When it comes to AI systems, the model's deployment environment and the formalization of the problem it's trying to solve may require some degree of explainability.[34] Low-risk or extensively studied and evaluated environments (e.g., spam filtering, optical character recognition) may not necessitate explanations from models. However, the need may arise when the problem's formalization is incomplete and a correct prediction is but a partial solution to the original problem.[34] In many instances, the loss function of the machine learning model doesn't cover additional constraints such as privacy, fairness, reliability (e.g., robustness to the avalanche or butterfly effect), or users' trust. Hence, explanations may be used to account for these constraints and meet the requirements of the deployment environment.

Given the variety of factors and the different legal contexts, some AI systems may be, and should be, required to provide explanations that are similar to those currently expected of human decision-makers and consistent with existing and upcoming standards specific to automated decision-making systems.[168] Designers of these AI systems (e.g., when engineering features and adjusting training datasets), internal users in an organization (e.g., when making a choice about the degree of trust they should place in an AI system), and customers (e.g., when understanding how they were affected by the decision of an AI system) can also find value in explaining the model's predictions, decisions, and actions.[145]

Beyond the context of decision-making, the demand for explainability is also driven by humanity's

search for scientific understanding.[34] Relying on large datasets and obscure learning methods to solve problems doesn't help us extract the additional knowledge captured by the model (e.g., causality). Explanations are the "show, don't tell" of AI systems and can help bridge the gap between the models' knowledge and ours by walking us through their learning and decision-making processes.

1.3 Intelligibility in Machine Learning

Having established the need for explainability in AI systems, we now turn to the technical considerations for extracting explanations. Our ability to generate explanations for the behavior of AI systems is constrained by the type of algorithms used: *transparent models* versus *black-box models*. This differentiation is mainly based on what has come to be called the model's complexity in the machine learning literature. This complexity is commonly and roughly evaluated in terms of the model's size.[3,58,150] Note that the exact estimation of a model's complexity may be difficult and the evaluation itself can be very subjective with respect to the end-user but we will discuss these issues in the next chapter when we cover the notion of interpretability.

Our ability to explain models is also constrained by the types of data we're dealing with. Forms of data such as tabular data, images, and texts are easily understandable while other forms of data such as sequence data, spatio-temporal data, and complex network data are hardly so.[58] Even for forms of data that are easily understandable (e.g., texts and images), their processing for the purposes of predictive models tends to require their transformation into vectors that are less understandable for humans so we may use equivalences, approximations or heuristics to allow this data to be used both by the AI system and for the model's explanations.[58]

1.3.1 Transparent Models

A small set of existing machine learning algorithms are considered inherently transparent. This set

includes decision trees, decision rules, and linear models.[24,42,73,97,150,175] These models are transparent because their internal components (e.g., weights, paths, or rules) are visible to an auditor and can be inspected to trace back the decision-making process. This set also includes K-Nearest Neighbors, general additive models, and Bayesian models,[11] but they are beyond the scope of this thesis.

Decision Trees

A decision tree model is a graph consisting of internal nodes representing tests on the features (e.g., boolean comparisons) and leaf nodes representing an outcome (i.e., class label).[58] The paths from the root to the leaves represent the classification rules and are linearizable into a set of decision rules with the if-then form (i.e., if x and y then z).[40,142,144] The separate analysis of each path from the leaf node to the root provides insights on composable "local knowledge."[58] Decision trees are also widely adopted for their graphical representation because the hierarchical position of the features in a tree provides immediate information about the most important attributes of a rule.[58] However, to remain interpretable, decision trees need to stay short because the number of terminal nodes increases exponentially with their depth. Moreover, decision trees are step-like prediction functions that implicitly categorise numeric features and are bad at describing linear relationships between features and outputs.[118] The depth of the tree is often adopted as a measure of the tree model's complexity.[150]

Decision Rules

Models based on decision rules are generalizations of those based on decision trees and map an observation to an outcome using association rules whose consequence is the outcome. The most common decision rules are if-then rules formed by conjunctions,[68] m-of-n rules where the consequence depends on the verification of m out of the n conditions,[122] and list of rules where the consequence depends on the first rule that's verified in a set of ordered rules.[182] In contrast to decision trees, de-

cision rules have a textual representation and some of them (e.g., ordered rule lists) can be harder to interpret than classical rules.[58,178] However, an attribute's relative importance can still be indicated through positional information and the study of single rules to analyze "local knowledge" remains possible.[58] Similar to decision trees, decision rules require their features to be categorical and are bad at describing linear relationships between features and outputs.[118]

Linear Models

Linear models are a class of models in which the output can be expressed as a linear combination of a series of features where no feature appears as a multiplier, divisor or exponent to any other feature. Linear regression and its different extensions are the most common linear models. A linear regression model predicts the output as a weighted sum of the features. The interpretation of a weight in the linear regression depends on whether it is associated with a numerical feature or a categorical feature. In the case of a numerical feature, increasing the feature leads to a proportional change in the estimated output by the associated weight. In the case of a categorical feature, changing the one category to another category changes the estimated output by the feature's weight with respect to the new category. Various statistics (e.g., t-statistic) and visualizations (e.g., effect plot) are available to measure the importance of a specific feature and other information relevant for comprehensibility.[118] The number of non-zero weights is often adopted as the measure of a linear model's complexity.[150]

1.3.2 Black Box Models

Deep learning algorithms like convolutional neural networks are considered black box models because their higher prediction accuracy comes at the expense of their transparency. Instead of relying on the developer's selection of features and data, these algorithms can process large amounts of information to learn by themselves which features are important. The internal components of

these algorithms (e.g., image pixel information, complex connections across several layers of neural networks) result from unexpected associations and are often uninterpretable to human users. To address this trade-off between accuracy and explainability, various explainabality methods, commonly referred to as post-hoc techniques, have been devised to turn black-box models into glass-box models. These techniques can be classified using two dimensions, whether they are model-specific or model agnostic and whether the explanations they provide are global in scope to the model or local in scope to a prediction. [145]

Model agnosticism is based on the idea that peaking into a model isn't necessary for the provision of explanations. The separation of explainability from the model allows the use of a variety of machine learning approaches to fulfill a single task. Moreover, this approach allows a single model to be explained with different types of explanations. [149] Model-agnostic approaches can generate explanations using different features than the underlying model. [149] Besides, the ability to explain a variety of models using the same techniques and representations is practical in real-world settings because it lowers the cost of switching from one model to the other when the system designer is comparing different approaches. [149] On the other hand, the dependence of model-specific techniques on the model to be interpreted may allow them to use this knowledge to generate more precise explanations.

The intuition behind changing the scope of the explanation is that approximating a black-box model by a transparent model in the neighborhood of the prediction we want to explain is easier than trying to approximate a model globally. [57] In fact, while global explanations can also be used to explain individual predictions, they can be less accurate than local explanations. However, for complex models, a user may find it hard to develop a global understanding of the model using local explainability methods because different local explanations from the same model can be inconsistent.

Model-Specific Local Explanations

Explainability techniques providing model-specific local explanations focus on explaining the

model's decision for a specific instance. In the example of decision trees, a transparent model, a model-specific local explanation would be equivalent to finding the tree path that led to the outcome for the specific instance. For black-box deep image classification convolutional networks, one approach consists in computing the gradient of the class score with respect to the input image using back-propagation techniques and generating image-specific class saliency visualisations. [161] These saliency maps are topographical representations of an image's influential regions.

Model-Specific Global Explanations

Explainability techniques providing model-specific global explanations incorporate constraints into the structure of the model. These constraints can be semantic meaningfulness constraints such as limits on the abstractions extracted from the data or interpretability constraints such as limits on the inputs' number of features. [145] For example, in the context of tree-based opaque machine learning models such as random forests and gradient boosted trees, TreeExplainer computes optimal local explanations for a specific model, extends local explanations to directly capture feature interactions using game theory, and provides insights on the model's global structure with some local faithfulness. [106]

Model-Agnostic Local Explanations

Explainability techniques providing model-agnostic local explanations generate explanations for a specific instance or for its vicinity. [150,164] For example, the Local Interpretable Model-Agnostic Explanations (LIME) technique perturbs the input around its neighborhood by changing comprehensible components (e.g., words or image regions), weights the model's predictions on these perturbed data points by their proximity to the original instance, and learns a transparent linear model on these intput-output associations. [150] Another famous technique is the Shapley additive explanations (SHAP) technique that turns feature values of a data instance into players in a coalition and uses the optimal

Shapley Values from coalitional game theory to assign each feature an importance value for a particular prediction. [105]

Model-Agnostic Global Explanations

Explainability techniques providing model-agnostic global explanations use input-output associations from a black-box model to develop a surrogate white-box model that fully approximates it. [145] For example, the Model Agnostic Globally Interpretable Explanations (MAGIX) approach repeatedly uses LIME for each instance in the training set and uses a genetic algorithm to evolve this set of locally important conditions at the global level. [141]

"[We] make the world smart so that we can be dumb in peace"

Andy Clark

2

The Obstacles in the Way of Interpretability

In the first chapter, we've established the need for explainable AI (XAI) systems and provided a short introduction to transparent models and explainability methods. In this second chapter, we will focus on interpretability and the meanings explanations lead to. Specifically, we will look at the factors that are relevant for a human-computer interaction with an explainable AI system. In Section 2.1, we will situate our focus within the *Doshi-Velez and Kim*'s taxonomy of evaluation approaches for a model's comprehensibility.[33] In Section 2.2, we will argue that interpretability is not a monolithic

concept and can't be captured by a single measure because it is heavily dependent on the context and the targeted audience. In Section 2.3, we will discuss several results from cognitive science that have interesting implications for interpretability and may explain our findings in Section 2.2. In Section 2.4, we will look at how interpretability tools can confuse a user's mental model of the AI system under investigation. Finally, in Section 2.5, we will argue that interpretability cannot be unidirectional and needs to be viewed as an interactive process between the user and the interpretability tools standing for the AI system.

2.1 Evaluation of Comprehensibility

To evaluate a model's comprehensibility, *Doshi-Velez and Kim* (2017) provide a taxonomy for the different approaches: application-grounded, human-grounded, and functionally-grounded.[33] The first, the application-grounded evaluation, is "an evaluation approach for interpretability" where the quality of an explanation is evaluated "in the context of its end-task, such as whether it results in better identification of errors, new facts, or less discrimination."[33] The second, the human-grounded evaluation, is an evaluation approach that's based on "more general notions of the quality of an explanation" such as comprehensibility under severe time constraints and relies on "abstract tasks" that maintain "the essence of the target application" and "in which other factors such as the overall task complexity can be controlled."[33] The third and last one, the functionally-grounded evaluation, abstracts away the results of a model's human-grounded evaluation and focuses on improving "explanation quality" with respect to, for example, prediction performance.[33]

According to these definitions, functionally-grounded evaluations restrict themselves to evaluating our notion of explainability while both application-grounded evaluations and human-grounded evaluations are focused on evaluating our notion of interpretability (See Introduction for these notions' definitions). Given our own focus on interpretability, this chapter restricts itself to human-

grounded evaluations and application-grounded evaluations. We cover some aspects of the human-computer interaction with an explainable AI (XAI) system that should be taken into account and inform these evaluations. We unpack abstractions such as "the context of the end-task" and "the essence of the target application" to make visible how these aspects play an important role in any assessment of interpretability.

2.2 Breaking Down the "Monolithic" Interpretability

In this section, we argue that interpretability is not a monolithic concept capturable by a single measure. Unlike the notion of explainability, interpretability depends heavily on the context, the assigned task, and the targeted audience and meaningful explanations can only meet our interpretability objectives if these dependencies are taken into account.

2.2.1 The Myth of The Single Measure

A model's size may be a good proxy for a model's interpretability,[3,95] but it is not the single measure. Our tendency to think that shorter explanations or simpler models perform best is an interpretation of Occam's razor principle. Attributed to the philosopher and theologian William of Ockham, this principle is often used as an inductive bias in the machine learning algorithms' selection of hypotheses to address the problem of running into an unlimited number of hypotheses.[143] However, Occam's razor was shown to be lacking in several contexts. For example, in the context of scientific theories, judgment of simplicity should not be made "solely on the linguistic form of the theory"[137] and should also account for semantic and pragmatic simplicity.[77,136,137] Similarly, a model's size is a syntactical aspect and accounts for none of the model's semantic[42] or pragmatic aspects.[88] Several studies agree with this assessment.[5,42,96] In a study by *Lavrac* (1999) on the importance of interpretability in certain medical applications informing medical decisions, medical experts were found to prefer larger

trees over shorter ones *at the expense of the models' accuracy* because the latter failed to meet their expectations about the required "sensitivity and specificity of the induced descriptions."[96] In an investigation by *Narayanan et al.* (2018), explanation complexity was shown to hurt a user's efficiency and satisfaction but not necessarily their accuracy.[124] The work of *Forough et al.* (2018) shows that the number of input features doesn't necessarily help people build better mental models.[138]

2.2.2 Interpretability Is Context-Dependent

The interpretability of different models and explanations can vary depending on the context and the generalizability of any evaluation of interpretability is constrained by its context. In fact, if evaluations of interpretability don't take the context into consideration, their results can seem contradictory when taken at face value. For example, the studies of *Subramanian et al.* (1992) and *Huysmans et al.* (2011) had different contexts, and these contexts led to opposite conclusions about whether decision trees or decision tables are more interpretable.[73,165] In the study of *Subramanian et al.* (1992) where 67 non-expert users were asked to interpret decision trees and decision tables to make investment decisions, the study concluded that trees are more interpretable than decision tables for the understanding of conditional logic.[165] In the study of *Huysmans et al.* (2011) where 51 non-expert users answered questions about a credit-scoring classification model in different representations, the experiment's results show that decision tables perform better than decision trees and decision rules when testing for the end users' accuracy, response time, and answer confidence.[73] The users' post-test voting is also consistent with these results and reveals their clear preference for decision tables (62.7%) over decision trees and decision rules in terms of ease of use.[73] Instead of questioning these studies for having different conclusions, it makes more sense to consider their results within their context. In fact, while explanations of the same type can be compared without any further context as part of a functionally-grounded evaluation,[154] explanations of different types (e.g., decision trees vs. decision tables, saliency maps vs. text captions) need to be evaluated within a context to account for, among other things, the

nature of the task assigned and the audience.[25]

2.2.3 User-Centered Explanations

As we've previously defined it, interpretability is the extent to which the explanations of a model's inner workings are comprehensible to a user agent. This comprehensibility can be tied to subjective factors. For example, to meet users' interpretability needs, models and their explanations need to earn the audience's acceptance, trust, and reliance. Unsurprisingly, the audience's acceptance of an explanation is affected by subjective preferences that can dictate the style of the explanations as much as the meaningfulness of their content.

Depending on the audience, explanations may have to look the part to meet the audience's interpretability needs.[2] For example, in an application in the Earth Sciences, Schwabacher and Langley (2001) found that the form of the learned models should match the form that is customarily used in the relevant literature to facilitate their acceptance in the corresponding scientific community.[95] In their work, they had to rely on a process model, stated in terms of differential equations and not just graphs, because their choice of algorithms for "aiding [Earth] scientists' understanding of data" was constrained by the field's "common formalism for representing knowledge."[95] These scientists couldn't "communicate their results" to their colleagues otherwise.[95] If *Schwabacher and Langley* were dealing with a different audience, differential equations would probably not be considered the most interpretable form for the explanations. In a different example from the medical domain, previous studies found that medical experts can be mentally opposed to over-simplistic explanations of complex relations.[36,96] In fact, they may choose complex models over simpler models that they deem "unnatural," even when this complexity comes at the expense of accuracy.[96] In other words, the expectations of the users can play a role in their acceptance of the explanations regardless of the explanations' ground truth. This conception of what a "natural" explanation looks like is similar to the agreed-upon "formalism for representing knowledge" that *Schwabacher and Langley* encountered within the com-

munity of Earth scientists. For medical experts, Bayesian classifiers with conditional probabilities for how much each feature contributes to a diagnosis and prognosis were found to be "natural" and satisfactory because they think that summing information gains in this manner is closer to how they'd diagnose patients and makes the most out of all the available information.[96]

Even within restricted scientific communities, there are individual differences in demands for interpretability that require tailored explanations. To better understand why data scientists need interpretability, *Hohman et al.* (2019) conducted a user study with 12 professional data scientists and found that reasons to interpret models included generating a hypothesis about the data and model, gaining insights into large datasets, and improving models with a better understanding of the underlying characteristics.[66] For each of these use cases, they found that explanations need to be tailored for the specific need.[66] Interestingly, the data scientists also recognized different scenarios in which they could use explanations to communicate what features were most predictive to stakeholders looking to deploy a model and acknowledged that "different audiences require different explanations" that balance succinctness and completeness.[66]

2.3 Cognitive Biases Are The Third Party In Human-Computer Interaction

In this section, we attempt to explain some of the results in Section 2.2 by drawing on existing research from cognitive science and education theory. Despite the benefits of cognitive biases and heuristics, we show that cognitive errors play an important role in any human-computer interaction and need to be taken into consideration. To avoid these pitfalls, researchers of other decision-making contexts have taken an interest in metacognitive strategies and we suggest that their insights should inform our recommendations for explainable AI (XAI) systems.

2.3.1 Human Rationality Borders Irrationality

In *Models of Man* (1957), *Herbet A. Simon* proposed the concept of bounded rationality to challenge rational choice theory and account for people's "limits in formulating and solving complex problems and in processing (receiving, storing, retrieving, transmitting) information."[159] To account for the unrealisticness of perfect rational decisions, bounded rationality assumes that human beings are rationally bounded, motivationally limited, and cognitively biased. Bounded rationality rethinks the norms based on optimization, utilities, and probabilities and focuses on studying the "actual behavior of minds and institutions."[50] Introduced in the 1970s, the term "cognitive bias" describes "people's systematic but purportedly flawed patterns of responses to judgment and decision problems."[179] Common to all human beings, these systematic biases and heuristics are a consequence of our "cognitive limitations, motivational factors, and/or adaptations to natural environments."[179] Given our restricted working memory and our brainpower's limitations, boundedly rational agents make decisions based on heuristics instead of optimization despite their inherent faultiness and reductionism. For example, even though chess and tic-tac-toe are both finite games with perfect information, fully rational behavior in chess is a lot harder than in tic-tac-toe because of the binding constraints of our mental capacities.[14]

2.3.2 The Silver Lining of Cognitive Biases

We often think of these biases as a shortcoming of human judgment but they actually play an important role in our day-to-day decision-making.[51] Some of our cognitive biases effectively support concept acquisition.[167] For example, our mutually exclusive bias, our tendency to infer "if not p then not q" after being convinced that "if p then q," was found to promote vocabulary growth in children.[109,115,167] Other advantages of cognitive biases include speeding up scrutiny to improve target detection in uncertain situations, supporting swift choice-making for practical short-term plans, al-

lowing the creation of fairly stable but imperfect categories to navigate the otherwise intractable world, and motivating the completion of problem-solving.[170] The impressive effectiveness of these biases have motivated research in introducing these cognitive biases into our models as inductive biases to reproduce this human-level concept learning.[62,157]

2.3.3 Pitting Interpretability Against Our Cognitive Biases

Given their importance in building our judgments, cognitive biases also affect human interpretations of AI systems and cognitive science should inform how explanations are presented. For example, decades of research by *Tversky and Kahneman* have established that judging the "human-perceived plausibility of hypotheses" should take into account that "similarity is more accessible than probability, that changes are more accessible than absolute values, that averages are more accessible than sums, and that the accessibility of a rule of logic or statistics can be temporarily increased by a reminder."[84] In this subsection, we attempt to explain some of the users' "irrational" behaviors through our cognitive biases.

Users' expectations towards the system's explanations may be influenced by our recognition heuristic. Our recognition heuristic is our tendency to infer that a recognized object rather than an unrecognized object has the higher value with respect to some criterion we cannot directly evaluate.[48,49,50] This heuristic may explain why *Schwabacher and Langley* needed their model to follow their audience's "common formalisms" of knowledge representation to ensure that it is accepted by the community of Earth scientists.[95] In general, accounting for the recognition heuristic may help secure the audience's acceptance.

Some users' preference for deeper trees and longer rules at the expense of accuracy may be explained by our representativeness heuristic. This heuristic is our tendency to make judgements based on similarity because we find a thing more likely when it is representative and similar to our existing preconceptions.[172] A common manifestation of this heuristic is the conjunction fallacy describing

how "a conjunction can be [found to be] more representative than one of its constituents."[44,172] A famous example of this fallacy is the Linda Problem: participants who are given the information that Linda is an outspoken woman, who majored in philosophy and was deeply concerned with issues of discrimination and social justice, overwhelmingly choose the statement "Linda is a bank teller and is active in the feminist movement" over the statement "Linda is a bank teller" when asked to pick the more probable one.[173] Given that decision rules and decisions trees can take the form of conjunctions, the users' interpretation of these models may be prone to the conjunction fallacy which may explain some users' preference for deeper trees and longer rules at the expense of accuracy. Hence, our representativeness heuristic can affect our interpretation of explanations.

Our confirmation bias may lead us to overtrust an AI system and its explanations. Our confirmation bias is our tendency to "[seek or interpret] evidence in ways that are partial to existing beliefs, expectations, or a hypothesis in hands."[128] In a study by *Lakkaraju et al.* (2020), they manage to elicit these expectations and take advantage of this bias to fool participants into trusting an untrustworthy black-box model.[94] These experiments show that users' confirmation bias can lead them to overtrust explanations that include desired features they think should be relevant and/or omit prohibited features they think shouldn't be relevant.[94] To meet our interpretability goals, it is important to take this confirmation bias into account in order to avoid inadvertently fooling our users into trusting our explanations

When "the available amount [of information...] makes [it] confusing and dysfunctional,"[134] information overload can affect our cognition and come in the way of our interpretability goals. The phenomenon of information overload has been observed to reduce the subjective choice accuracy in different settings.[59,76,108] Several studies in cognitive science describe cognition as a property of the whole system within which we function.[27,55,69,70,71,72,130,183] As a result, our cognitive threshold is influenced by changes in environmental properties (e.g., an increase in the amount of information presented).[98] Several studies found that choice accuracy decreases with the number of options or the

number of attributes[134] and increases with primed attribute information.[108,155] In the context of interpretability, *Poursabzi-Sangdeh et al.* (2021) found that information overload can lead increased transparency to hamper people's ability to detect a model's sizable mistakes or unusual inputs.[138] The effect of the amount of information on a user's behavior is often described by an inverse U-shaped curve.[126,134] However, we also believe that information design, the way information is presented, can affect our cognitive threshold. As we've seen earlier, presenting some information in the form of a decision table can be easier to process than presenting the same information in the form of a decision tree.[73] In other words, evaluations of the effect of information on users need to take into account three dimensions: users' accuracy, the information design, and the amount of information. Regardless, given that the amount of information presented affects our cognition, it should be taken into account in the assessment of interpretability.

Our literature review shows that insights from cognitive science that should inform the design decisions behind explainable AI (XAI) systems are underexplored. In this section, we show how some of these results can explain the variations observed in the users' interpretability. We use some of these connections to motivate our design recommendation in Chapter 3. However, beyond the objectives of this thesis, we hope that this discussion can motivate further conversations about the implications of our bounded rationality for our interpretability needs.

2.3.4 A Way Out of our Cognitive Pitfalls

Given the implications of bounded rationality for our decision-making, we investigated some metacognitive strategies for the activation of our more "deliberative mode of thinking" and the mitigation of the cognitive errors that may arise during our human-computer interactions.[85]. Introduced by Flavell in the 1970s, metacognition is best understood as "thinking about thinking,"[38,117] and encompasses an awareness of the requirements of the learning process, a recognition of the limitations of memory, an ability to appreciate perspective, and a capacity for self-critque.[31] Common strategies for

promoting metacognition include self-questioning,[180] thinking aloud,[146] and making graphic representations of one's thoughts and knowledge.[162] The development of metacognitive skills in the realm of explainable AI systems has not been explored before and the closest work we could find is in the contexts of flight crew decision-making and clinical decision-making.

In the context of flight crew decision-making, *Orasanu-Engel and Mosier* (2019) devise approaches to train crews in making effective decisions by looking at the factors that contributed to errors in former aircraft accidents.[131] In this context, they found that cognitive errors can cause faulty situation assessments (e.g., a misinterpretation of the available cues), a susceptibility to automation bias (e.g., relying on pattern recognition instead of more vigilant information search), faulty selections of course actions, or inadequate risk assessments (e.g., underestimation of the likelihood of possible consequences).[131] They consider metacognition to be the "most trainable decision-supporting skill" and argue for approaches that make explicit the metacognitive processes of "questioning [the crew's] interpretation of the situation and simulating the consequences of their decision[s]."[131] They found that the explicit provocation of the decision-maker's questioning and simulating can help in the development of metacognition.[131]

In the context of clinical decision-making, *Croskerry* (2003) presents three cognitive forcing strategies to avoid the cognitive errors that "underlie most diagnostic errors that are made [...] in the emergency department."[31] These strategies are "formal cognitive debiasing approach[es] to deal with [...] pitfalls in clinical reasoning" and "prevent clinicians from pursuing a pattern recognition path that typically will lead to error."[31] *Croskerry* (2003) differentiates between his proposed forcing strategies that require the clinician's conscious application of metacognition and *Lewis and Norman*'s "baked-in" forcing functions that are built into the system's design to minimize or avoid errors.[31,98] While *Lewis and Norman* force these strategies on the users, Croskerry's cognitive forcing strategies require the decision-maker's conscious choice to partake in metacognition. Unfortunately, experimental evidence suggests that the application and retention of such conscious forcing strategies is poor in the

clinical decision-making setting.[156] Hence, when viable, the "baked-in" strategies of *Lewis and Norman* may be more promising than *Croskerry*'s opt-in strategies for the development of metacognition.

Insights from these decision-making contexts can inform our designs of explainable AI (XAI) systems because these decision-making contexts have a lot in common with the interpretability context. In the flight crew decision-making context, flight crews are making decisions based on cues provided by the equipment at their disposal. Similarly, users of interpretability tools are making sense of the explainable cues provided by the AI system. Furthermore, the limited effectiveness of opt-in cognitive forcing strategies in clinical decision-making seems to indicate that "baked-in" forcing strategies are generally more effective. Finally, both studies conclude that the explicitation of the questioniong of a decision helps the development of metacognitive skills. In the next chapter, we use these insights to support the argument for conversational explainable AI (XAI) systems.

2.4 The Veil of Interpretability Tools

In this section, we argue that the availability of the interpretability tools themselves can have unexpected side effects on users' trust that go against their interpretability goals. In fact, when relying on interpretability tools to understand an AI system, the users' mental model of the system is veiled by their mental model of the interpretability tools. The user's mental model of a tool is their representation of the relationships between its various parts and is informed by their interaction with it.[129] As users rely on these tools to audit or understand an AI system, they often fail to disassociate their mental model of the tools and their mental model of the underlying system. In a study by *Kaur et al.* (2020) in which data scientists were given access to a training dataset and some interpretability tools and were asked to assess the reliability of a model, the authors found that many of these participants took the provided explanations at face value and used their mere existence to convince themselves that the underlying models were reliable and ready for deployment.[86] Given that any attempt to inter-

pret an AI system happens behind the veil of interpretability tools, a mismatch between the designers' conceptual model, the representation of the tool that the designers intended for the users to understand, [129] and the user's mental model can lead a user's trust of the interpretability tool to sway their trust of the underlying AI system. Interestingly, in *Kaur et al.*'s study, the participants' perception of the tools and the authority they gave to them weren't even informed by these tools' capabilities and were shrouded by their social context. [86] For example, the novelty and public availability of the tools used, InterpretML and the SHAP Python package, led some of the data scientists to trust them without fully understanding them. [86] One of the participants' comment is very telling: "I guess this is a publicly available tool... must be doing something right. I think it makes sense." [86] Given that the user's failure to build an accurate mental model of the interpretability tool can be as detrimental as their failure to build an accurate mental model of the underlying AI system, a good interpretability tool needs to be able to rein in these expectations.

2.5 Interpretability Can Only Be Interactive

In this section, we will argue that any explaining of an AI system is a sort of communication between the AI system represented by the interpretability tool on one hand and the end-user on the other. Moreover, to meet their end of this communication, interpretability tools need to be interactive, built with the view that a meaningful human-computer interaction is a bidirectional, dynamic process. Just as the assumptive error of treating human communication as a static entity rather than a dynamic process has hampered its investigation for a long time, [10] the view of interpretability as a static and unidirectional transaction will only hamper the evaluation of its effectiveness.

The explaining of an AI system is a communication between the AI system represented by the interpretability tool on one hand and the end-user on the other. Communication stands for "those acts in which meaning develops within human beings." [10] *Barnlund* (2008)'s Transactional Model of

Communication is based on the idea that meaning isn't "received" but is instead "invented."[10] The model states that the sender and the receiver invent this shared meaning together and are both responsible for the outcome of the communication.[10] The goal of interpretability is to help users understand the underlying AI system and give meaning to its structure, parts, and inner-workings from the explanations provided. Hence, the process of explaining an AI system is a sort of communication. In this communication, the sender is the interpretability tool standing for the AI system and the receiver is the end-user. In this communication, the end-user can fully understand an AI system's behavior once they have a "working model" of the system that somewhat accurately represents "what causes [this behavior], what results from it, how to influence, control, initiate, or prevent it, how it relates to other states of affairs or how it resembles them, how to predict its onset and course, what its internal or underlying "structure" is."[81] A "working model" doesn't need to perfectly represent what it's meant to model to be useful; it only needs to accurately model the behavior the end-user seeks to understand.

An explanation is "a blueprint for the construction of [an AI system's] working model"[81] and the end-user's final interpretation is the refined mental or working model. To explain the formation of these working models, *Kenneth Craik* (1943) suggests that reasoning consists in the manipulation of working models through three distinct processes: (1) A translation of an interaction into an internal representation in symbols (e.g., words, numbers); (2) The derivation of other symbols from the internalized symbols through inference; and (3) A recognition of the correspondence between these symbols and the observed external process.[12,29] As we add information about the world, we are going through these 3 steps repeatedly to refine our "working model" of it to our satisfaction. *Weld and Bansal* (2018) refer to this process of refining our "working model" as drilling down and following up and many members of the HCI community argue that it should be part of any explainable AI (XAI) system.[86,177]

For a user to trust their interpretation of an explanation and the explanation itself, interpretability tools need to give them this ability to drill down. As we saw before, a satisfactory "blueprint" for one

end-user may be unsatisfactory for another because of a variety of reasons (See Section 2.2). Moreover, in the context of inscrutable models, explanation methods tend to map a transparent model to a black-box model so they rely on approximations and necessarily lose some information that may conceal important details.[177] Drilling down by seeking more targeted information out of the interpretability tool allows the user to leave this communication with an interpretation and a "working model" to their satisfaction. An example of this ability to drill down is a suggested solution to the variety of preferences for models' sizes that consists in allowing users of explainable AI (XAI) systems to provide size constraints on the explanations in order to balance the model's comprehensibility and the users' preferences and needs all while avoiding a one-size-fits-all solution.[95,174] To adapt to the end user's goals and needs in this manner, XAI systems need to be interactive, bidirectional, and dynamic.

"The word 'communication' stands for those acts in which meaning develops within human beings [...] It arises out of the need to reduce uncertainty, to act effectively, to defend or strengthen the ego. Its aim is to increase the number and consistency of meanings within the limits set by attitude and action patterns that have proven successful in the past, emerging needs and drives, and the demands of the physical and social setting of the moment."

Dean C. Barnlund

3

Conversational Explanation Systems

In the first chapter, we've established the need for explainable AI (XAI) systems and provided a summary of the state-of-the-art in providing explanations from transparent and black-box models. In the second chapter, we've looked at the obstacles in the way of making meaning of these explanations. In this third and final chapter, we defend a specific approach to designing explainable AI (XAI) systems and present our design for a study to assess the effectiveness of this recommendation.

In the second chapter, we've established that the human-computer interaction between a user and

the interpretability tool standing in for the AI system needs to be thought of as a communication. This communication needs to support the refining of the user's "working model" of the AI system by giving them the ability to drill-down and follow-up. The right interpretability tool also needs to have the ability to adapt to its users' needs and goals by providing a variety of explanations within the constrains of the users' personal preferences and biases. Ideally, these tools should be designed with the common cognitive errors in mind to "bake-in" forcing strategies that support metacognitive processes.

In this chapter, we defend the design of explainable AI (XAI) systems as conversational explanation systems with those obstacles in mind. In Section 3.1, we present the work of *Weld and Bansal* (2018) and flesh out the arguments for their design recommendation.[177] In Section 3.2, we will look at how this design recommendation can account for the obstacles discussed in the previous chapter and potentially mitigate some of them. Finally, in Section 3.3, we will present a proposal for a user study that's motivated by previous work with Elizabeth Hu and Nari Johnson. The proposal in its current form is the fruit of an on-going collaboration between Elizabeth Hu and I under the supervision of Elena Glassman.

3.1 Explaining by Conversing

Weld and Bansal (2018) were the first to recommend the design of explainable AI (XAI) systems in the form of an interactive, conversational system.[177] The vision they sketched out for building interactive explanation systems allows these systems to adapt to the user's needs and support different follow-up and drill-down actions.[177] Their design recommendations were informed by the prior work of *Lim and Dey* (2009) in ubiquitous computing on the types of information demands users have and their implications in context-aware applications (i.e., applications that ground their behavior on information about the state of people, places, and objects relevant to the users and their activities).[101]

3.1.1 Intelligibility Types

By taking a usability-centric approach, *Lim and Dey* (2009) elicited a set of intelligibility types that users of context-aware applications may be interested in asking about and grounded these types in the users' underlying reasoning processes. [101] This suite of intelligibility types was developed from a set of common colloquial questions that participants in their user study asked. [101] These different types of explanations can support a variety of goals including filtering for causes, generalizing and learning a mental model, and predicting and controlling for a system's behavior. [102]

To meet one's interpretability goals in the more general applications of AI models, *Lim et al.* (2019) selected and refined 7 intelligibility types: "Inputs" explanations, "What Output" explanations, "Certainty" explanations, "Why" explanations, "Why Not" explanations, "What If" explanations, and "When" explanations. [102] "Inputs" explanations inform users about the input values that the application is reasoning from for the current instance. [102] "What Output" explanations inform users about the current prediction and what possible output label the application can produce. [102] "Certainty" explanations inform users about how (un)certain the application is of the output value produced. [102] "Why" explanations inform users why the application derived its output value from the current input values. [102] For example, these explanations may inform the user of the model's most influential features. "Why Not" explanations provide information about why an alternative outcome was not produced. [102] For example, these explanations may help the user infer what changes in an input could lead to the desired output. "What If" explanations allow users to simulate what the application will do given a user-set input values or changes. [102] For example, these explanations would allow the user to query the model with their own set of input data that's chosen to meet the user's needs. Finally, "When" explanations inform the user about the cases in which a user-set outcome would happen. [102] For example, these explanations can provide the user with examples of inputs that lead to a specific, queried outcome.

3.1.2 Interactive Explanatory Dialog

The proposal of *Weld and Bansal* (2018) is consistent with *Lim, Yang, and Wang*'s suite of intelligibility types for explainability and envisions different follow-up and drill-down actions.[177] These actions include, among others, asking for more detail about a decision, asking for a decision's rationale, or perturbing the input example.[177] When asking for more detail about a decision, the user may restrict the model to a subregion of the feature space such that the interactive explanation system may rely on local explanations with higher local accuracy.[177] When asking for a decision's rationale, the interactive explanation system may use nearest-neighbor methods to inform the user of the labeled training examples that were most influential in the underlying model's decision.[177] When perturbing the input example, the user can simulate and test their hypotheses about the model's inner workings to decide for themselves how accurate their "working model" for the AI system is.[177]

By supporting this variety of follow-up and drill-down actions, these interactive explanation systems are better-equipped to adapt to the users' needs and backgrounds. In fact, when checking a model's behavior, the user study of *Lim et al.* (2019) already shows that users tend to exploit different strategies and use different intelligibility queries for the same interpretability objectives.[102] To augment this capacity even further, *Weld and Bansal* (2018) suggest that explanation systems could build explicit models of users' knowledge and misconceptions.[177] However, if these interpretability tools are expected to explain any arbitrary black-box model, they do recognize that existing approaches from intelligent tutoring systems (ITSs) for building such models about the users may need to be expanded on further.[177] We note that this specific avenue of exploration may benefit from the flexibility of model agnostic approaches.

3.2 Defending the Call for Conversational XAI Systems

In this section, we provide our own arguments for Weld and Bansal's proposal by reconsidering

the obstacles laid out in the previous chapter in the light of a conversational explanation system where users can ask questions inspired by these intelligibility types to the chatbot of the interpretability tool about the underlying AI system.

3.2.1 Polylithic Interpretability

An interactive, conversational explanation system is particularly well-suited for offering the maleability required to adapt to the targeted audience and the end user's goals, needs, and context. In the previous chapter, we've established that interpretability is not a monolithic concept but is dependent on the context and the targeted audience. To avoid limiting our understanding of the end users' interpretability needs by the model's size, an interactive, conversational explanation system can easily allow for user-provided size constraints on the explanations to avoid a one-size-fits-all solution. These constraints can be changed dynamically during the interaction to provide the user with more flexibility. Moreover, given that preferences for explanations vary depending on the context, conversational explanation systems can naturally let the users drive the interaction to suit their goals and needs rather than demanding that the system infers them. Recent work by *Chen et al.* (2020) on context-aware explainable conversational recommendation models that incorporate user feedback show promising results in improving recommendation accuracy, meeting users' explainability needs, and working within user-set constraints. [26]

3.2.2 Interpretability with "baked-in" Metacognition

When interacting with a conversational explanation system, users will be forced to converse in writing with the chatbot of the interpretability tool. This requirement will act as a "baked-in" forcing strategy because extensive research indicates that the process of writing helps the development of metacognitive skills. Across education theory research, the writing process is often regarded as similar

to the thinking process and provides a way for people to automatically engage in metacognition and become active learners.[8,46] With regard to the relationship between writing and mathematical problem solving, the use of writing facilitates people's visualization of mathematical thinking in words and helps them describe each step of the problem solving process.[13,15,41,135,140] Writing helps people reflect and think critically about content by creating a personal transaction through which they take ownership of learning and build meaning.[46,111] It also steers them to self-question, infer from prior knowledge, and use their imaginations in order to produce novel thoughts and insights.[125,46] Writing also helps people develop reasoning skills and provides them with a way to organize and analyze the material they have read.[37,46,78,93] All these benefits of writing will help support the users' interpretability goals as they are conversing with a conversational explanation system and working on refining their working models of the underlying AI system.

When interacting with a conversational explanation system, users will be expected to write out questions about the model and its decisions that the interpretability tool will attempt to answer through a variety of explanations. This process of question formulation is also a metacognitive process in its own right. For example, to improve students' reading comprehension, the Question Formulation Technique (QFT) is a popular forcing strategy that stimulates students' awareness of their learning difficulties through question formulation.[89] Students use QFT as a self-monitoring technique to summarize a text by formulating questions and improve their reading comprehension as a result.[89] This strategy pushes people into assuming a more active role in the learning process.[92,132] By leaving some of the questions formulated open, this strategy further "[stimulates] student' curiosity and [invites] them to search further for the answer."[89] Interacting with a conversational explanation system will have the same effect as QFT and act as another "baked-in" forcing strategy. Users will find themselves required to assume the role of active learners, formulate questions about the underlying model, and follow-up on the information the chatbot provides.

Given the benefits of writing and formulating questions for developing metacognitive skills, we

hypothesize that an interactive conversational explanation system will act as a metacognitive process that's consistent with the interpretability goals of both the user and the designer. In the process of writing and formulating questions, the user will be forced to process the explanations they receive from the tool and figure out which insights they're missing to build a working model of the AI system that's accurate enough for their needs. Given that they are put in the position of an active learner, they will have to make sure to read and comprehend the explanations provided to generate new questions and investigate their hypotheses about the model's behavior. Previous research on troubleshooting computer systems shows that people tend to fixate, or rely on repetitions, rather than generate specific procedures based upon reasoning about the particular system.[147] By developing their metacognitive skills in this way, we hope to activate their more deliberative modes of thinking.

3.2.3 Unveiling the AI System

To separate the AI system from the interpretability tool, a conversational explanation system can help the user draw the line between the tool and the underlying AI system. This line can be drawn by using carefully crafted answers to the user's questions. For example, by referring to itself as a separate entity from the explained AI system, the chatbot can lower the risk of confusing or misleading the user into wrongly attributing errors to, or misplacing trust in, the wrong party. In a conversational setting, the explanation system can also provide background knowledge to better handle questions and provide complete answers.[45] Moreover, people are well-equipped to assess the limits and mistakes of a conversational agent when these chatbots fail to respond as expected or show limitations to the questions they can answer. In fact, previous studies have shown that users are able to gauge the chatbots from the quality of their interpretations of requests and advice,[39] and adapt by using more restricted vocabulary.[65] In contrast, as evidenced by the study of Karu et al. (2020), other forms of interpretability tools are harder to objectively and naturally assess.[86] In fact, Turing considered fooling human beings through conversational machines to be an interesting challenge for this exact reason.[171]

Note that future chatbots based on GPT-3 and their impact on users' expectations towards chatbots are considered to be beyond the scope of this thesis.

3.2.4 Bidirectional Interpretability

Conversational explanation systems are by definition interactive and bidirectional. Thy are built with a bidirectional user-model communication in mind and are motivated by the users' questions and needs that the work of *Lim et al.* (2019) on intelligibility types elicited. [101,102] To promote this bidirectionality, it is important to engage the user and make them comfortable with the interaction. For example, we believe that the tone used by the explanation system can help users build more accurate levels of trust towards it and the underlying AI system and weaken the impact of the social context associated with the tools on the user's assessment. This intuition stems from established research about error messages that we believe should inform the language used by conversational explainability chatbots to promote a question-friendly environment. A good example of this research is the work of *Shneiderman* (1980). *Shneiderman* (1980) shows that, when offensive, the format and tone of error messages can lead users, especially beginners, to attribute any ambiguity or misunderstanding to their own incompetence. [98,158] To remedy these problems, *Lewis and Norman* recommend adopting a user centered design and relying on a more apologetic tone to make the user more comfortable sharing the responsibility with the system. [98] In general, the language used by the conversational explanation system can play a big role in creating a safe environment that supports the user's interpretability goals and turn the interaction with interpretability tools and AI systems into a cooperative endeavor.

3.3 Our Proposal: A Wizard-of-Oz Experiment

In this section, we propose a Wizard-of-Oz study whose goal is to test some of these arguments. First, we define what we mean by a Wizard-of-Oz experiment. Second, we introduce the study we

designed with Elizabeth Hu and include information about the task description, the participants, the procedure, and the results' analysis. Finally, we present some preliminary results from earlier work with Nari Johnson and Elizabeth Hu.

3.3.1 Wizard-of-Oz Experiments

Pioneered by *Gould et al.* in 1983,[54] Wizard-of-Oz (WoZ) experiments in the field of human-computer interaction are commonly used to speed up the prototyping of costly systems or elicit people's requirements and expectations from futuristic systems. These WoZ experiments are research experiments in which participants are made to interact with a computer system that they believe to be autonomous but is actually operated or guided by a human being, the Wizard. While WoZ experiments are powerful tools for research, our work takes into consideration the recommendations of *Maulsby et al.* (1993) about the importance of limiting the Wizard's intelligence and freedom and basing their capabilities on formal models in order to ensure consistent interaction, honest simulation, and appropriate results.[110]

3.3.2 Randomized Experimental Design

Inspired by the proposal of *Weld and Bansal* (2018) for conversational explanation systems, Elizabeth Hu and I built off of our project with Nari Johnson and designed this Wizard-of-Oz experiment as a user study to:

- **(G1)** Assess participants' satisfaction with conversational XAI interfaces,
- **(G2)** Elicit implicit user expectations towards an explainability chatbot, and
- **(G3)** Investigate whether this design proposal can help develop users' metacognitive skills.

Task Description

Our experiment was inspired by *Weld and Bansal*'s example of an interactive explanatory dialog. We assigned participants the task of predicting the decision of a simulated image classification model of clownfish on 6 images. We choose image classification for our user study because of the task's general accessibility. [20]

For each image, participants are given the ground truth label for whether or not a clownfish is present in the image and are asked to predict whether the AI system will decide that there is a clownfish in the image or not (See Appendix A). To inform their predictions, participants are given access to explanations. Depending on the group they are assigned to (See Procedure), participants are either able to ask for these explanations themselves in a conversation with the chatbot by following a format provided by a question bank and inspired by the intelligibility types or they are given access to the explanatory answers of all the possible questions they could have asked conversationally at once. Both the AI system and the explanations are simulated. They are pre-scripted and inspired by current research in explanation methods. The same images, simulated model, and explanations are used across all participants and they were all made to be consistent with a distribution shift, a common AI failure mode resulting from lacking training data. In our case, we assumed that the AI system was trained with data that over-represented photos of both clownfish and anemone living symbiotically. This assumed distribution shift was meant to explain why 2 of the 6 images were misclassified by the simulated AI system as depicting clownfish for including anemones (i.e. false positives).

Participants

The recruited participants are undergraduate and graduate students from Harvard University. All participants are recruited through public university mailing lists after they are asked to fill out a demographic survey. The demographic survey is meant to verify their age and English fluency and

collect some data about their academic background. Participants are offered a $10 compensation for their time to incentivize their participation. This compensation was made possible by the support of the Glassman Lab at the Harvard John A. Paulson School of Engineering and Applied Sciences and the Harvard College Research Program (HCRP).

Procedure

The study is conducted online on Zoom and Slack. Recruited participants are divided randomly into two groups: the control group and the experimental group. Both groups are asked to agree to a consent form before they are led to watching a 3-minute primer YouTube video on key concepts and terms in Artificial Intelligence to provide all participants with the same baseline knowledge for the study. Once they are done watching the video, the study operator provides them with a brief description of the study without giving any hints about the simulated aspects of the experiment (See Appendix A).

In the experimental variation of the protocol, study participants are given a question bank inspired by the intelligibility types outlined above. We limit them to a few "Certainty" explanations (**C1, C2**), "Why" explanations (**W1, W2, W3**), and "What If" explanations (**WI1, WI2, WI3**) to provide enough variety, minimize the cognitive load, and simulate realistic capabilities. Below is a list of the questions that users can ask of the model:

- **C1**: How confident is the model in this prediction?
- **C2**: What is the accuracy of the model?
- **W1**: What regions of the image are most influential to the prediction?
- **W2**: What image features are most influential to the prediction?
- **W3**: Which training examples were most influential to the prediction?
- **WI1**: What happens when object [1-9] is removed from this image?

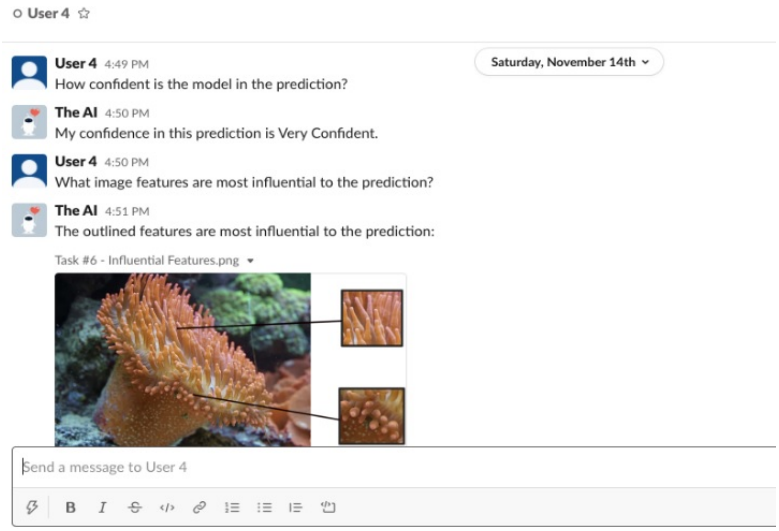

Figure 3.1: An example of an interactive explanatory dialog between a participant and the chabot.

- **WI2**: What happens when there is also a {user-specified} present in the photo?
- **WI3**: What happens if the color of object [1-9] is {user-specified}?

Once they are ready, study participants can freely converse with the simulated conversational explanation system on the Slack platform. We choose the Slack platform for its support of chatbots and people's familiarity with it. We implemented a rule-based chatbot that can recognize questions from the question bank and respond with our pre-scripted answers. The chatbot responds to questions one-by-one with images and/or text depending on the exact question and in the order they are asked (See Figure 3.1). If the participant's query isn't recognized, the chatbot responds with an apologetic message: "Sorry, {user's name}! I can't answer your question. Could you try a different question from the question bank?" Participants can ask as many questions as they want before sharing their prediction with the chatbot. The chatbot then responds with the model's actual decision: "Thank you for your response! Your guess was {right/wrong}. My actual prediction for this task was: {YES/NO}."

In the control variation of the protocol, study participants are given all the answers to all the

questions they could have asked conversationally in a file without conversing with the explanation interface. They are expected to read these explanations, comprehend them, and make a prediction about the model's decision. This file of questions and answers is the script we used to implement the chatbot and the only difference between the two settings is the conversational interactivity of the experimental variation. Given that the model and tasks are otherwise the same, this control group allows us to directly compare the effect of conversational explanation system to a setting that is non-interactive and isolate the effect of conversational interactivity.

Design and Analysis

After each of the 6 tasks, participants are asked to answer a short questionnaire that is meant to track the development of their trust-level throughout the study. This questionnaire includes the following four questions: "What do you think is the overall accuracy of the AI?", "To what extent do you believe you can trust the decisions the AI will make", "How would you rate the expected performance of the AI relative to you expected performance for clownfish identification?", "I feel that I understand how the AI works {much better, better, somewhat better, the same, somewhat worse, worse, much worse} than before interacting with the model." These questions are designed as multiple-choice questions to reduce the cognitive load on the user and take as little time as possible.

At the end of the study, each participant is also asked to fill out a longer questionnaire with more detailed questions. These questions are meant to collect data on the user's trust, mental model development, and their satisfaction with the conversational explanation system. This questionnaire includes the following questions:

- **Q1**: What do you think is the overall accuracy of the AI?
- **Q2**: How would you rate the expected performance of the AI relative to your expected performance for clownfish identification?

- **Q3**: To what extent do you believe you can trust the decisions the AI will make?
- **Q4**: I feel that I understand how the AI works {much better, better, somewhat better, the same, somewhat worse, worse, much worse} than before interacting with the model.
- **Q5**: Asking which question contributed most to your understanding of the AI? (please specify just 1 question or intelligibility type)
- **Q6**: How likely are you to deploy this AI for use in a practical environment for a clownfish-identifying task? The concept of deployment in data science refers to the application of an AI model for prediction using new data. Building an AI is generally not the end of the project. Deployment is the method by which you integrate an AI into an existing production environment to make practical business decisions based on data.
- **Q7**: In 1-2 sentences, please describe how you think the AI determines whether a clownfish is present or not (i.e. what information you used to predict the AI's decisions).
- **Q8**: What did you think of your interactions with the AI? Did you like the conversational format or would you have preferred all the answers to the Question Categories provided at once, without having to ask for them?

The answer format for each of these questions is as follows: percentage answer (**Q1**), 7-point Likert scale (**Q2, Q3, Q4, Q6**), and free-form response (**Q5, Q7, Q8**). Many of these questions are needed to compensate for the fact that we don't push participants to think aloud in order to avoid adding an external cognitive forcing strategy.

We also track the participant's performance on these tasks, including their accuracy and the time they spend per task. This data allows us to assess the effectiveness of these explanation systems in supporting the interpretability goals of users.

3.3.3 Results from Preliminary Experiments

Given that this study is still under review by the Committee on the Use of Human Subjects (CUHS) that serves as the Institutional Review Board for Harvard University, we haven't started running this study yet. However, for educational purposes and as part of a course, we did run the interactive version of this experiment a few times with Elizabeth Hu and Nari Johnson to improve our study design and collect some data to motivate the larger experiment. These experiments didn't rely on a chatbot but on a study operator that simulated the bot. We recruited 11 participants, 7 of them had some background in Computer Science or other adjacent fields.

While preliminary, our results were consistent with many of our predictions and hypotheses. In terms of trust, the results indicated that, on average, participants tended to somewhat distrust the model, as they should. This distrust may be attributed to identifying the failure mode within the AI system and corresponds to an accurate level of trust. When asked whether they would deploy the model for use in a practical environment, participants' average response was what they were somewhat likely to deploy it. In terms of satisfaction, most participants (8/11) enjoyed the conversational interaction and found it better suited for the investigation of an AI model than the static format we described to them. As one of the participants said, "I liked that it was open-ended because it allowed me to be more deductive about [...] reason[ing] with the AI [...] the fact that the questions were presented one after another made me feel like I had to ask every question very intentionally, on a path to discovery, rather than being given a cheat sheet and looking for patterns." Alluding to information overload, another participant commented that they "didn't have to deal with an influx of information all at once." In general, even though participants did spend more time on the tasks that corresponded to the false positives (See Task 3 and 6 in Figure 3.2), it is hard to assess the effectiveness of this design for the development of metacognitive skills without a control group so we couldn't really draw satisfactory conclusions from these experimental runs. However, these results do seem to indicate

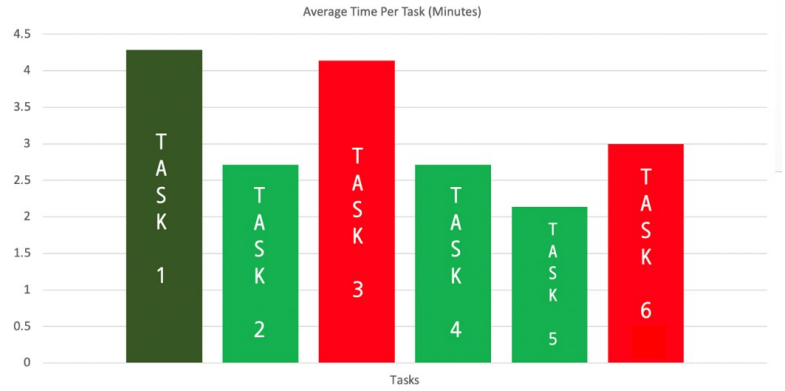

Figure 3.2: An example of an interactive explanatory dialog between a participant and the chabot.

that participants build an accurate level of trust towards the AI system and are generally satisfied with conversational XAI interfaces because they do support their interpretability goals.

Despite the general positive feedback, some participants did express some reservations about our simulated conversational explanation system. Participants questioned the realisticness of the use of this system in an industry setting where time and efficiency are highly valued. While participants took 3 minutes per task on average, they weren't sure that an actual machine learning practitioner with more complicated models and contexts would be willing to to go through the hassle of conversing before vetting a model for deployment. Similarly, many participants worried that, in the long run, they might get bored from the restrictive back-and-forth format. They suggested allowing them to ask multiple questions at once to reduce wait time. It would be interesting to see how far we can go along the spectrum between interactive explanation systems and static explanation systems before losing many of the benefits of interactivity (e.g., forcing strategies). Another interesting finding was that some users complained of the limitations of natural language and wished they could refer to objects and regions on the pictures by describing them or even pointing at them.

4

Conclusion

This thesis follows our personal journey from exploring why we need explainable AI systems to researching how to best meet users' interpretability needs. Artificial intelligence has become an inseparable part of our daily life. As long as we are willing to allow our technology to shape the world around us, we need to accept the idea that other fields' perspectives can direct our work. The same way isolationism from the rest of computer science stumped the growth of artificial intelligence in its infancy, the isolationism of computer science from other research areas will only stump the growth

of artificial intelligence in this critical stage of its development. This thesis draws on research from the fields of human-computer interaction, machine learning, public interest technology, cognitive science, and education theory to make the case for a specific research direction in explainable AI (XAI) systems. Beyond the technical difficulty of explainability, the quest for interpretability is inherently a human process strewn with cognitive, contextual, and practical considerations. To account for them, explainable AI (XAI) systems need to take into consideration users' needs and expectations. With some of these needs and expectations in mind, we argued that the design recommendation of *Weld and Bansal* (2018) for conversational explanation systems is not only a natural research direction but also a promising proposal given the obstacles in the way of human-interpretable AI.[177]

As we wait on the review of the university's Institutional Review Board, we can only look forward to study the results of our proposed experiment and explore several future directions. One direction consists in expanding on naive conversational explanation systems to better meet the elicited expectations of users (e.g., providing them with the ability to point to objects in the picture and refer to them in conversation). We also believe that actually implementing our simulated chatbot will bring up interesting technical problems for software engineers, statisticians, and UX researchers. The challenge of building on the explanation methods considered in Section 1.3 of Chapter 1 and meeting the intelligibility expectations outlined in Section 3.2 of Chapter 3 is in and of itself an interesting avenue of research. As we scale our experiment, it would be interesting to investigate how our results and the the expectations of users vary across a variety of subgroups (e.g., experts vs. non-experts, students vs. professionals). Given our preliminary results, we expect to find interesting differences across these subgroups in the way they interact with the model.

A

User Study Instructions

All the information in this Appendix is from my on-going collaboration with Elizabeth Hu and inspired by previous work with Nari Johnson and Elizabeth Hu. Both have been supervised by Elena Glassman.

Instructions

You will be presented with a series of pictures that either have or do not have a clownfish.

A clownfish is a type of *fish* that is an overall *yellow*, *orange*, or *blackish color*, and many show distinctive *white stripes*. *Clownfish* are typically found in sheltered reefs or lagoons and live in a symbiotic relationship with surrounding sea *anemones*, a type of *marine animal*.

We also provide you with a sentence confirming whether or not a clownfish is in the picture, if in any case you are unsure about the contents of the picture.

A particular Artificial Intelligence (AI) system does not know whether or not a clownfish is present in each picture. The AI tries to evaluate whether or not a clownfish is in a picture.

You will be able to interact with the AI on Slack by asking the AI questions. Your questions must be exactly one of the listed **Question Bank** questions below. There is no limit to the number of questions you may ask, and you will be able to refer back to the Question Categories anytime you wish. We expect that you will take fewer than 5 minutes to ask questions per task.

Based on your interaction with the AI, your goal is to **predict the AI's decision for each picture.** There is only one AI in the experiment: the same AI makes predictions for all of the pictures. After you predict the AI's decision for a picture, the AI's true decision will be revealed to you.

Please read the Question Bank on the next page. It outlines the exact questions and formats you can ask of the AI.

Figure A.1: We provide these introductory instructions to the participants in the conversational setting of our user study.

TASK #1

When you're ready to start the study, please navigate to Browse Slack → Apps → Click "The AI" → Messages. Please message *Task 1* under Messages and begin the task.

The AI must decide: Is there a clownfish in this photo?

Fact: There is NOT a clownfish in this photo.

After interacting with the AI, your goal is to answer the Assessment Question:

Assessment Question: What will the AI decide?
Options:

- YES = there is a clownfish in this photo.
- NO = there is not a clownfish in this photo

You may now ask the AI questions by sending messages. Use the AI agent's responses to guide your Assessment Question answer. Each question you ask must exactly adhere to 1 of the listed questions in the **Question Bank**. There is no limit to the number of questions you may ask.

Figure A.2: This is the first task assigned to the participants of our study.

References

[1] Abbasova, V. (2020). US Analyzes Azerbaijani Military Tactics During Karabakh War. *Caspian News*.

[2] Achinstein, P. (1983). *The nature of explanation*. New York: Oxford University Press.

[3] Afrabandpey, H., Peltola, T., Piironen, J., Vehtari, A., & Kaski, S. (2020). A decision-theoretic approach for model interpretability in bayesian framework. *Machine learning*, 109(9-10), 1855.

[4] Agar, J. (2020). What is science for? The Lighthill report on artificial intelligence reinterpreted. *The British Journal for the History of Science*, 53(3), 289–310.

[5] Allahyari, H. & Lavesson, N. (2011). *User-oriented Assessment of Classification Model Understandability*.

[6] Allen-Ebrahimian, B. (2019). Exposed: China's Operating Manuals for Mass Internment and Arrest by Algorithm.

[7] Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

[8] Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The Effects of School-Based Writing-to-Learn Interventions on Academic Achievement: A Meta-Analysis. *Review of Educational Research*, 74(1), 29–58.

[9] Banko, M. & Brill, E. (2001). Scaling to very very large corpora for natural language disambiguation. In *Proceedings of ACL 2001*.

[10] Barnlund, D. C. (2008). A transactional model of communication. In *communication theory* (pp. 47–57). Routledge, 2 edition.

[11] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58, 82–115.

[12] Beaubien, R. (2013). The Nature of Explanation.

[13] Bell, E. S. & Bell, R. N. (1985). Writing and Mathematical Problem Solving: Arguments in Favor of Synthesis. *School Science and Mathematics*, 85(3), 210–221.

[14] Bendor, J. (2015). Bounded rationality. In J. D. Wright (Ed.), *International Encyclopedia of the Social Behavioral Sciences (Second Edition)* (pp. 773–776). Oxford: Elsevier, second edition edition.

[15] Berkenkotter, C. (1982). Writing and problem solving. In *Language connections : writing and reading across the curriculum*. Urbana, Ill.: National Council of Teachers of English.

[16] Bibal, A. & Frénay, B. (2016). Interpretability of machine learning models and representations: an introduction.

[17] Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car.

[18] Boole, G. (1992). George boole and the mathematics of logic. 1847. *M.D. computing*, 9(3), 165.

[19] Bromley, A. G. (1998). Charles babbage's analytical engine, 1838. *IEEE Annals of the History of Computing*, 20(4), 29–45.

[20] Buccinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20 (pp. 454–464). New York, NY, USA: Association for Computing Machinery.

[21] Buchanan, B. G. & Mitchell, T. M. (1978). Model-directed learning of production rules. In D. A. Waterman & F. Hayes-Roth (Eds.), *Pattern-Directed Inference Systems* (pp. 297–312). Academic Press.

[22] Buolamwini, J. & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler & C. Wilson (Eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research* (pp. 77–91). New York, NY, USA: PMLR.

[23] Carlini, N. & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods.

[24] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on knowledge discovery and data mining*, KDD '15 (pp. 1721–1730).: ACM.

[25] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R. M., Kelley, T. D., Braines, D., Sensoy, M., Willis, C. J., & Gurram, P. (2017). Interpretability of deep learning models: A survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)* (pp. 1–6).: IEEE.

[26] Chen, Z., Wang, X., Xie, X., Parsana, M., Soni, A., Ao, X., & Chen, E. (2020). Towards Explainable Conversational Recommendation. volume 3 (pp. 2994–3000).

[27] Clark, A. (1997). *Being there : putting brain, body, and world together again*. Cambridge, Mass.: MIT Press.

[28] Cohen, P. R. (2017). *Empirical methods for artificial intelligence.* MIT Press. OCLC: 1032721220.

[29] Craik, K. J. W. (2013). *Nature of Explanation.* Cambridge University Press. OCLC: 1122719896.

[30] Crevier, D. (1993). *AI: the tumultuous history of the search for artificial intelligence*. New York, NY: Basic Books.

[31] Croskerry, P. (2003). Cognitive forcing strategies in clinical decisionmaking. *Annals of emergency medicine*, 41(1), 110–120.

[32] Cross, S. E. & Walker, E. (1994). Dart: applying knowledge-based planning and scheduling to crisis action planning.

[33] Doshi-Velez, F. & Kim, B. (2017a). Towards a rigorous science of interpretable machine learning.

[34] Doshi-Velez, F. & Kim, B. (2017b). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*. arXiv: 1702.08608.

[35] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2019). Accountability of AI Under the Law: The Role of Explanation. *arXiv:1711.01134 [cs, stat]*. arXiv: 1711.01134.

[36] Elomaa, T. (1994). In defense of c4.5: Notes on learning one-level decision trees. In *Machine Learning Proceedings 1994* (pp. 62–69). Elsevier Inc.

[37] Fellows, N. J. (1994). A window into thinking: Using student writing to understand conceptual change in science learning. *Journal of Research in Science Teaching*, 31(9), 985–1001.

[38] Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, 34(10), 906–911.

[39] Følstad, A., Nordheim, C. B., & Bjørkli, C. A. (2018). What makes users trust a chatbot for customer service? an exploratory interview study. In S. S. Bodrunova (Ed.), *Internet Science* (pp. 194–208). Cham: Springer International Publishing.

[40] Frank, E. & Witten, I. H. (1998). *Generating accurate rule sets without global optimization*, volume 98/2. of *Working paper*. Hamilton, N.Z: Dept. of Computer Science, University of Waikato.

[41] Frank K. Lester, Jr., Joe Garofalo, & Diana Lambdin Kroll (1989). *The Role of Metacognition in Mathematical Problem Solving: A Study of Two Grade Seven Classes*. Technical report.

[42] Freitas, A. (2014). Comprehensible classification models: a position paper. *SIGKDD explorations*, 15(1), 1–10.

[43] Friedrich, V. S. & Jolmes, J. (2020). Amazon: Der Vorgesetzte sieht alles. *tagesschau*.

[44] Fuernkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine learning*, 109(4), 853–898.

[45] Galitsky, B. (2020). Conversational Explainability. In B. Galitsky (Ed.), *Artificial Intelligence for Customer Relationship Management: Keeping Customers Informed*, Human–Computer Interaction Series (pp. 415–445). Cham: Springer International Publishing.

[46] Gammill, D. M. (2006). Learning the write way. *The Reading teacher*, 59(8), 754–762.

[47] Garg, A. X., Adhikari, N. K. J., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., Sam, J., & Haynes, R. B. (2005). Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review. *JAMA*, 293(10), 1223.

[48] Gigerenzer, G. (1999). *Simple heuristics that make us smart*. Evolution and cognition. New York: Oxford University Press.

[49] Gigerenzer, G. (2015). *Simply rational : decision making in the real world*. Oxford series in evolution and cognition. Oxford ; New York: Oxford University Press.

[50] Gigerenzer, G. & Selten, R., Eds. (2001). MIT Press.

[51] Goldstein, D. G. & Gigerenzer, G. (1999). The recognition heuristic: How ignorance makes us smart. In G. Gigerenzer, P. M. Todd, & A. R. Group (Eds.), *Simple heuristics that make us smart* (pp. 37–58). New York: Oxford University Press.

[52] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples.

[53] Goodman, D. & Keene, R. (1997). *Man versus machine: Kasparov versus Deep Blue*. Cambridge, Mass: H3 Publications.

[54] Gould, J. D., Conti, J., & Hovanyecz, T. (1983). Composing letters with a simulated listening typewriter. *Communications of the ACM*, 26(4), 295–308.

[55] Greeno, J. G. & Moore, J. L. (1993). Situativity and symbols: Response to vera and simon. *Cognitive science*, 17(1), 49–59.

[56] Guedes, N. (2020). É impossível fiscalizar a obrigação de ter a aplicação StayAway Covid. *TSF Rádio Notícias*.

[57] Guestrin, Sameer Singh, C. M. T. R. (2016). Local Interpretable Model-Agnostic Explanations (LIME): An Introduction.

[58] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Pedreschi, D., & Giannotti, F. (2018). A Survey Of Methods For Explaining Black Box Models. *arXiv:1802.01933 [cs]*. arXiv: 1802.01933.

[59] Hahn, M., Lawson, R., & Lee, Y. G. (1992). The effects of time pressure and information load on decision quality. *Psychology & Marketing*, 9(5), 365–378.

[60] Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, 24(2), 8–12.

[61] Harwell, D. (2020). Algorithms are deciding who gets the first vaccines. Should we trust them? *Washington Post*.

[62] Hattori, M. & Oaksford, M. (2007). Adaptive non-interventional heuristics for covariation detection in causal induction: Model comparison and rational analysis. *Cognitive science*, 31(5), 765–814.

[63] Hays, J. & Efros, A. A. (2007). Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH 2007)*, 26(3).

[64] Hern, A. (2020). Microsoft apologises for feature criticised as workplace surveillance. *The Guardian*.

[65] Hill, J., Randolph Ford, W., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior*, 49, 245–250.

[66] Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. (2019). Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*, CHI'19 (pp. 1–13).: ACM.

[67] Hu, X., Zhao, Y., Deng, L., Liang, L., Zuo, P., Ye, J., Lin, Y., & Xie, Y. (2020). Practical attacks on deep neural networks by memory trojaning. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, (pp. 1–1).

[68] Huang, L., Chen, H., Wang, X., & Chen, G. (2000). A fast algorithm for mining association rules. *Journal of computer science and technology*, 15(6), 619–624.

[69] Hutchins, E. (1995a). *Cognition in the wild*. Cambridge, Mass.: MIT Press.

[70] Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive science*, 19(3), 265–288.

[71] Hutchins, E. & Klausen, T. (2008). Distributed cognition in an airline cockpit. In Y. Engeström & D. Middleton (Eds.), *Methods in Language and Social Interaction*, volume 4 of *SAGE Benchmarks in Social Research Methods* (pp. IV340). London: SAGE Publications Ltd.

[72] Hutchins, E. & Palen, L. (1997). Constructing Meaning from Space, Gesture, and Speech. In L. B. Resnick, R. Säljö, C. Pontecorvo, & B. Burge (Eds.), *Discourse, Tools and Reasoning: Essays on Situated Cognition*, NATO ASI Series (pp. 23–40). Berlin, Heidelberg: Springer.

[73] Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems*, 51(1), 141–154.

[74] Interel the Group Secretariat (2019). *Report from the Gambling Related Harm All-Party Parliamentary Group: Online Gambling HarmInquiry*. Technical report, Gambling Related Harm All Party Parliamentary Group.

[75] Jackson, P. (1999). *Introduction to expert systems*. International computer science series. Harlow, England ; Reading, Mass: Addison-Wesley, 3rd ed edition.

[76] Jacoby, J., Speller, D. E., & Berning, C. K. (1974). Brand Choice Behavior as a Function of Information Load: Replication and Extension. *Journal of Consumer Research*, 1(1), 33–42.

[77] Jahn, M., Herman, D., & Ryan, M.-L. (2010). *Routledge Encyclopedia of Narrative Theory*. Taylor and Francis.

[78] Jeanne Swafford, J. K. B. (2000). Instructional Strategies for Promoting Conceptual Change: Supporting Middle School Students. *Reading & Writing Quarterly*, 16(2), 139–161.

[79] Jha, S., Raman, V., Pinto, A., Sahai, T., & Francis, M. (2017). On learning sparse boolean formulae for explaining ai decisions. In *NASA Formal Methods*, Lecture Notes in Computer Science (pp. 99–114). Cham: Springer International Publishing.

[80] Jia, R. & Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems.

[81] Johnson-Laird, P. N. P. N. (1983). *Mental models : towards a cognitive science of language, inference, and consciousness*. Cognitive science series ; 6. Cambridge, Mass.: Harvard University Press.

[82] Juneja, P. & Mitra, T. (2021). Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation.

[83] Jónsson, A. K., Morris, P. H., Muscettola, N., Rajan, K., & Smith, B. (2000). Planning in interplanetary space: theory and practice. In *Proceedings of the Fifth International Conference on Artificial Intelligence Planning Systems*, AIPS'00 (pp. 177–186). Breckenridge, CO, USA: AAAI Press.

[84] Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *The American psychologist*, 58(9), 697–720.

[85] Kahneman, D. (2013). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux, 1st pbk. ed edition. OCLC: ocn834531418.

[86] Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Wortman Vaughan, J. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *CHI 2020*. CHI 2020 Honorable Mention Award.

[87] Klabbers, J. (2006). The right to be taken seriously: Self-determination in international law. *Human rights quarterly*, 28(1), 186–206.

[88] Kliegr, T. (2017). Effect of cognitive biases on human understanding of rule-based machine learning models.

[89] Koch, A. & Eckstein, S. G. (1991). Improvement of reading comprehension of physics texts by students' question formulation. *International Journal of Science Education*, 13(4), 473–485.

[90] Koebler, J. (2020). Detroit Police Chief: Facial Recognition Software Misidentifies 96% of the Time.

[91] Koh, P. W. & Liang, P. (2017). Understanding black-box predictions via influence functions.

[92] Kourilsky, M., Esfandiari, M., & Wittrock, M. C. (1996). Generative teaching and personality characteristics of student teachers. *Teaching and Teacher Education*, 12(4), 355–363.

[93] Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, 77(3), 319–337.

[94] Lakkaraju, H. & Bastani, O. (2020). "how do i fool you?": Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on ai, ethics, and society*, AIES '20 (pp. 79–85).: ACM.

[95] Langley, P., Norvig, P., & Schwabacher, M. (2001). Discovering communicable scientific knowledge from spatio-temporal data.

[96] Lavrač, N. (1999). Selected techniques for data mining in medicine. *Artificial intelligence in medicine*, 16(1), 3–23.

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

[98] Lewis, C. & Norman, D. A. (1995). Designing for error. In R. M. Baecker, J. Grudin, W. A. Buxton, & S. Greenberg (Eds.), *Readings in Human–Computer Interaction*, Interactive Technologies (pp. 686–697). Morgan Kaufmann.

[99] Lewis, H., Ed. (2021). *Ideas that created the future : classic papers of computer science*. Cambridge, Massachusetts: The MIT Press.

[100] Liang, B., Li, H., Su, M., Bian, P., Li, X., & Shi, W. (2018). Deep text classification can be fooled. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*.

[101] Lim, B. Y. & Dey, A. K. (2009). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing*, UbiComp '09 (pp. 195–204). Orlando, Florida, USA: Association for Computing Machinery.

[102] Lim, B. Y., Yang, Q., Abdul, A., & Wang, D. (2019). Why these explanations? selecting intelligibility types for explanation goal.

[103] Lipton, Z. C. (2016). The mythos of model interpretability.

[104] Longman, M. & Schulte, B. (2020). Why Today's Shopping Sucks. *Washington Monthly*, January/February/March 2020.

[105] Lundberg, S. & Lee, S.-I. (2017). A unified approach to interpreting model predictions.

[106] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2019). Explainable AI for Trees: From Local Explanations to Global Understanding. *arXiv:1905.04610 [cs, stat]*. arXiv: 1905.04610.

[107] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

[108] Malhotra, N. K., Jain, A. K., & Lagakos, S. W. (1982). The information overload controversy: An alternative viewpoint. *Journal of Marketing*, 46(2), 27–37.

[109] Markman, E. M. & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2), 121–157.

[110] Maulsby, D., Greenberg, S., & Mander, R. (1993). Prototyping an intelligent agent through Wizard of Oz. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, CHI '93 (pp. 277–284). Amsterdam, The Netherlands: Association for Computing Machinery.

[111] Mayher, J. S., Lester, N. B., & Pradl, G. M. (1983). *Learning to write/writing to learn*. Upper Montclair, N.J: Boynton/Cook.

[112] McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (2006). A proposal for the dartmouth summer research project on artificial intelligence: August 31, 1955. *The AI magazine*, 27(4), 12.

[113] McDermott, J. (1982). R1: A rule-based configurer of computer systems. *Artificial Intelligence*, 19(1), 39–88.

[114] McEneney, M. F. & Kaufmann, K. F. (2005). Implementing the fact act: Self-executing provisions. *The Business lawyer*, 60(2), 737–747.

[115] Merriman, W. E. (1989). *The mutual exclusivity bias in children's word learning*. Monographs of the Society for Research in Child Development ; v. 54, no. 3-4, serial no. 220. Chicago, IL: Society for Research in Child Development.

[116] Messier, W. F. J., Quilliam, W. C., Hirst, D. E., & Craig, D. (1992). The effect of accountability on judgment: Development of hypotheses for auditing; discussions; reply. *Auditing : a journal of practice and theory*, 11, 123.

[117] Metcalfe, J. & Shimamura, A. P., Eds. (1994). *Metacognition: Knowing about Knowing*.

[118] Molnar, C. (2019). *Interpretable Machine Learning*.

[119] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations.

[120] Morvan, V.-X. (2020). À Cannes, des caméras détectent les personnes masquées. *Le Figaro*.

[121] Mozur, P., Zhong, R., & Krolik, A. (2020). In Coronavirus Fight, China Gives Citizens a Color Code, With Red Flags. *The New York Times*.

[122] Murphy, P. M. & Pazzani, M. J. (1991). Id2-of-3: Constructive induction of m-of-n concepts for discriminators in decision trees. In *Machine Learning Proceedings 1991* (pp. 183–187). Elsevier Inc.

[123] Naiara Bellio (2020). Spanish police plan to extend use of its lie-detector while efficacy is unclear.

[124] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., & Doshi-Velez, F. (2018). How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation.

[125] National Writing Project & Nagin, C. (2003). *Because writing matters : improving student writing in our schools*. Jossey-Bass education series. San Francisco: Jossey-Bass.

[126] Nematzadeh, A., Ciampaglia, G. L., Ahn, Y.-Y., & Flammini, A. (2019). Information overload in group communication: from conversation to cacophony in the Twitch chat. *Royal Society Open Science*, 6(10), 191412.

[127] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images.

[128] Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2), 175–220.

[129] Norman, D. A. (1987). *Some Observations on Mental Models*, (pp. 241–244). Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA.

[130] Norman, D. A. (1993). *Things that make us smart : defending human attributes in the age of the machine*. Reading, Mass.: Addison-Wesley Pub. Co.

[131] Orasanu-Engel, J. & Mosier, K. L. (2019). Chapter 5 - flight crew decision-making. In B. G. Kanki, J. Anca, & T. R. Chidester (Eds.), *Crew Resource Management (Third Edition)* (pp. 139–183). Academic Press, third edition edition.

[132] Osborne, R. J. & Wittrock, M. C. (1983). Learning science: A generative process. *Science Education*, 67(4), 489–508.

[133] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2015). The limitations of deep learning in adversarial settings.

[134] Pilli, L. E. & Mazzon, J. A. (2016). Information overload, choice deferral, and moderating role of need for cognition: Empirical evidence. *Revista de administração (São Paulo)*, 51(1), 36–55.

[135] Poh, B. L. G. & Sam, L. C. (2015). The Impact of Mathematical Writing on Students' Metacognition in Applied Algebra Test. *International Journal of Social Science Studies*, 4(1), 18–27.

[136] Popper, K. R. K. R. (1959). *The logic of scientific discovery*. New York: Basic Books.

[137] Post, H. R. (1960). Simplicity in scientific theories. *The British journal for the philosophy of science*, 11(41), 32–41.

[138] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., & Wallach, H. (2018). Manipulating and measuring model interpretability.

[139] Preece, A., Harborne, D., Braines, D., Tomsett, R., & Chakraborty, S. (2018). Stakeholders in Explainable AI. *arXiv:1810.00184 [cs]*. arXiv: 1810.00184.

[140] Pugalee, D. K. (2001). Writing, Mathematics, and Metacognition: Looking for Connections Through Students' Work in Mathematical Problem Solving. *School Science and Mathematics*, 101(5), 236–245.

[141] Puri, N., Gupta, P., Agarwal, P., Verma, S., & Krishnamurthy, B. (2018). MAGIX: Model Agnostic Globally Interpretable Explanations. *arXiv:1706.07160 [cs]*. arXiv: 1706.07160.

[142] Quinlan, J. (1999). Simplifying decision trees. *International journal of human-computer studies*, 51(2), 497–510.

[143] Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1), 81–106.

[144] Quinlan, J. R. (1987). Generating production rules from decision trees. In *Proceedings of the 10th international joint conference on Artificial intelligence - Volume 1*, IJCAI'87 (pp. 304–307). Milan, Italy: Morgan Kaufmann Publishers Inc.

[145] Rai, A. (2020). Explainable ai: from black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141.

[146] Raihan, D. (2011). 'think-aloud' techniques used in metacognition to enhance self-regulated learning. 25, 1738–2246.

[147] Rasmussen, J. & Jensen, A. (1974). Mental Procedures in Real-Life Tasks: A Case Study of Electronic Trouble Shooting. *Ergonomics*, 17(3), 293–307.

[148] Rees, M. (2020). Censure de photos, drones, reconnaissance faciale... déluge sécuritaire à l'Assemblée. *Next Inpact*.

[149] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016a). Model-Agnostic Interpretability of Machine Learning. *arXiv:1606.05386 [cs, stat]*. arXiv: 1606.05386.

[150] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016b). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*. arXiv: 1602.04938.

[151] Rosenblueth, A., Wiener, N., & Bigelow, J. (1943). Behavior, purpose and teleology. *Philosophy of science*, 10(1), 18–24.

[152] Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Computational models of cognition and perception. Cambridge, Mass: MIT Press.

[153] Russell, S. & Norvig, P. (2013). *Artificial Intelligence: A Modern Approach*. Pearson custom library. Pearson Education UK.

[154] Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Muller, K.-R. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE transaction on neural networks and learning systems*, 28(11), 2660–2673.

[155] Scammon, D. L. (1977). "Information Load" and Consumers. *Journal of Consumer Research*, 4(3), 148–155.

[156] Sherbino, J., Dore, K. L., Siu, E., & Norman, G. R. (2011). The effectiveness of cognitive forcing strategies to decrease diagnostic error: An exploratory study. *Teaching and learning in medicine*, 23(1), 78–84.

[157] Shinohara, S., Taguchi, R., Katsurada, K., & Nitta, T. (2007). A model of belief formation based on causality and application to n-armed bandit problem. *Transactions of the Japanese Society for Artificial Intelligence*, 22(1), 58–68.

[158] Shneiderman, B. (1980). *Software psychology : human factors in computer and information systems*. Winthrop computer systems series. Cambridge, Mass.: Winthrop Publishers.

[159] Simon, H. A. (1957). *Models of man: social and rational; mathematical essays on rational human behavior in a social setting*. New York: Wiley.

[160] Simon, H. A. & Newell, A. (1958). Heuristic Problem Solving: The Next Advance in Operations Research. *Operations Research*, 6(1), 1–10.

[161] Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*. arXiv: 1312.6034.

[162] Soedjoko, E., Suyitno, H., & Rochmad (2019). Representation of students metacognition in constructing of graphics. *Journal of physics. Conference series*, 1321(2), 22091.

[163] Strauss, D. A. (1989). Discriminatory intent and the taming of brown. *The University of Chicago law review*, 56(3), 935–1015.

[164] Strumbelj, E. & Kononenko, I. (2010). An Efficient Explanation of Individual Classifications using Game Theory. *The Journal of Machine Learning Research*, 11, 1–18.

[165] Subramanian, G., Nosek, J., Raghunathan, S., & Kanitkar, S. (1992). A comparison of the decision table and tree. *Communications of the ACM*, 35(1), 89–94.

[166] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks.

[167] Taniguchi, H., Sato, H., & Shirakawa, T. (2018). A machine learning model with human cognitive biases capable of learning from small and biased datasets. *Scientific reports*, 8(1), 7397–13.

[168] The Government of Canada (2019). Directive on Automated Decision-Making.

[169] Thrun, S., Montemerlo, M., Dahlkamp, H., Stavens, D., Aron, A., Diebel, J., Fong, P., Gale, J., Halpenny, M., Hoffmann, G., Lau, K., Oakley, C., Palatucci, M., Pratt, V., Stang, P., Strohband, S., Dupont, C., Jendrossek, L.-E., Koelen, C., Markey, C., Rummel, C., van Niekerk, J., Jensen, E., Alessandrini, P., Bradski, G., Davies, B., Ettinger, S., Kaehler, A., Nefian, A., & Mahoney, P. (2007). *Stanley: The Robot That Won the DARPA Grand Challenge*, (pp. 1–43). Springer Berlin Heidelberg: Berlin, Heidelberg.

[170] Tobena, A., Marks, I., & Dar, R. (1999). Advantages of bias and prejudice: an exploration of their neurocognitive templates. *Neuroscience & Biobehavioral Reviews*, 23(7), 1047–1058.

[171] Turing, A. M. (1995). *Computing Machinery and Intelligence*, (pp. 23–46). American Association for Artificial Intelligence: USA.

[172] Tversky, A. & Kahneman, D. (1981). Evidential impact of base rates. In *Judgment under uncertainty : heuristics and biases*. Cambridge ; New York: Cambridge University Press.

[173] Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4), 293–315.

[174] Van Assche, A. & Blockeel, H. (2017). Seeing the forest through the trees: Learning a comprehensible model from a first order ensemble. In *Inductive Logic Programming*, Lecture Notes in Computer Science (pp. 269–279). Berlin, Heidelberg: Springer Berlin Heidelberg.

[175] Wang, F. & Rudin, C. (2015). Causal falling rule lists.

[176] Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1), 36–45.

[177] Weld, D. S. & Bansal, G. (2018). The challenge of crafting intelligible intelligence.

[178] Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *The Artificial intelligence review*, 11(1), 273–314.

[179] Wilke, A. & Mata, R. (2012). In V. S. Ramachandran (Ed.), *Encyclopedia of Human Behavior*. San Diego: Elsevier Science Technology.

[180] Williamson, R. (1996). Self-Questioning — An Aid to Metacognition. *Reading Horizons: A Journal of Literacy and Language Arts*, 37(1).

[181] Winograd, T. (1971). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.

[182] Yoon, J. & Kim, D.-W. (2012). Classification based on predictive association rules of incomplete data. *IEICE transactions on information and systems*, E95.D(5), 1531–1535.

[183] Zhang, J. & Norman, D. A. (1994). Representations in distributed cognitive tasks. *Cognitive science*, 18(1), 87–122.