Why these Explanations?
Selecting Intelligibility Types for Explanation Goals

Brian Y. Lim
National University of Singapore
Singapore
brianlim@comp.nus.edu.sg

Qian Yang
Carnegie Mellon University
Pittsburgh, PA, USA
yangqian@cmu.edu

Ashraf Abdul, Danding Wang
National University of Singapore
Singapore
{ashrafabdul,wangdanding}@u.nus.edu

ABSTRACT

The increasing ubiquity of artificial intelligence (AI) has spurred the development of explainable AI (XAI) to make AI more understandable. Even as novel algorithms for explanation are being developed, researchers have called for more human interpretability. While empirical user studies can be conducted to evaluate explanation effectiveness, it remains unclear why specific explanations are helpful for understanding. We leverage a recently developed conceptual framework for user-centric reasoned XAI that draws from foundational concepts in philosophy, cognitive psychology, and AI to identify pathways for how user reasoning drives XAI needs. We identified targeted strategies for applying XAI facilities to improve understanding, trust and decision performance. We discuss how our framework can be extended and applied to other domains that need user-centric XAI. This position paper seeks to promote the design of XAI features based on human reasoning needs.

CCS CONCEPTS

• Human-centered computing ~ Human computer interaction

KEYWORDS

Intelligibility; Explanations; Explainable artificial intelligence; Decision making

ACM Reference format:

Brian Y. Lim, Qian Yang, Ashraf Abdul and Danding Wang. 2019. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In *Joint Proceedings of the ACM IUI 2019 Workshops, Los Angeles, USA, March 20, 2019, 7 pages.* https://doi.org/10.1145/1234567890

1 Introduction

The recent success of artificial intelligence (AI) is driving its prevalence and pervasiveness in many domains of decision making from supporting healthcare intervention decisions to informing criminal justice. However, to ensure that we understand how these models and algorithms work, and to better control them, these models need to be explainable. As a result, explainable AI research has been burgeoning with many algorithmic approaches being developed to explain AI and many HCI driven empirical studies to understand the impact of these explanations. We refer the interested reader to several literature reviews [1, 5, 10, 43].

To help end users to understand, trust, and effectively manage their intelligent partners, HCI and AI research have produced many user-centered, innovative algorithm visualizations, interfaces and toolkits (e.g., [7, 20, 25, 36]). To make sense of the variety of explanations, several explanation frameworks have been proposed for knowledge-based systems [10], recommender systems [12], case-based reasoning [39], intelligent decision aids [40], tutoring systems [10], intelligible context-aware systems [24], etc. These frameworks are mostly taxonomic or driven by clearly defined principles (e.g. [21]). In this work, we aim to identify theories in human thinking that drives the needs for different types of explanations.

Indeed, some work has drawn from more formal theories. Recent writings by Miller, Hoffman and Klein discussed relevant theories from philosophy, cognitive psychology, social science, and AI to inform the design of eXplainable AI (XAI) [13, 14, 15, 19, 31]. Miller noted that much of XAI research tended to use the researchers' intuition of what constitutes a "good" explanation. He argued that to make XAI usable, it is important to draw from social sciences. Hoffman et al. [13, 14, 15] and Klein [19] summarized several theoretical foundations of how people formulate and accept explanations, empirically identified several purposes and patterns for causal reasoning, and proposed ways that users can generate self-explanations to answer contrastive questions. However, it is not clear how best to operationalize this rich body of work in the context of XAI-based decision support systems for specific user reasoning goals. Hence, adding on to this line of inquiry, we have recently proposed a theory-driven, user-centric XAI framework that connects XAI explanation features to underlying reasoning processes that users have for explanations [42]. Drawing on this framework, XAI researchers and designers can identify pathways along which human cognitive patterns drives needs for building XAI. By articulating a detailed design space of technical features of XAI and user requirements of human reasoning, we intend that our framework will help

IUI Workshops'19, March 20, 2019, Los Angeles, USA.

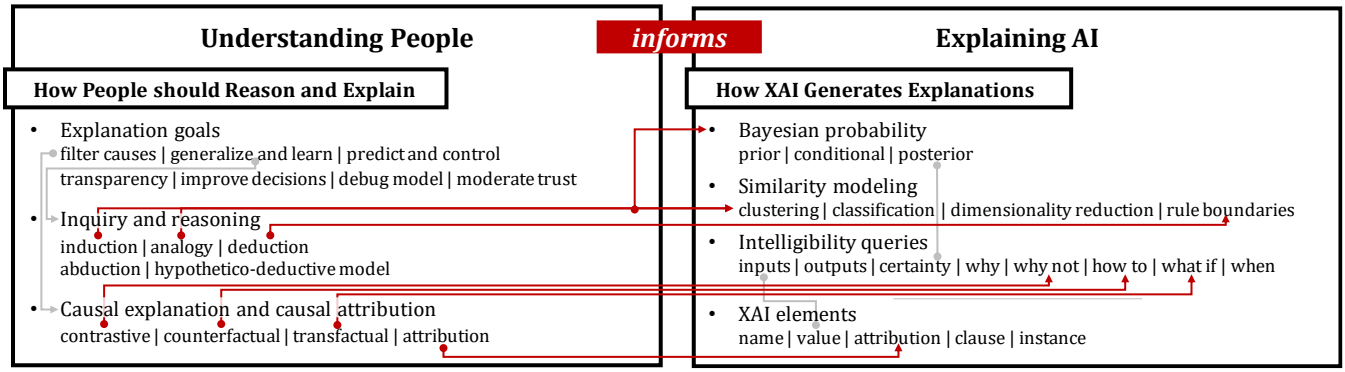

Figure 1. Partial Conceptual framework for Reasoned Explanations (of [42]) that describes how human reasoning processes (left) informs XAI techniques (right). Points describe different theories of reasoning, XAI techniques, and strategies for designing XAI. Arrows indicate pathway connections: red arrows for how theories of human reasoning inform XAI features, and grey for inter-relations between different reasoning processes and associations between XAI features. Only some example pathways are shown. For example, to find the cause of an application behavior, user could seek a contrastive explanation of counterfactuals to filter causes (grey arrow); this can be supported with why not and how to explanations, respectively (red arrows). To help users generalize and learn about the application behavior, we should support reasoning processes (grey arrow) of induction, analogy and deduction by highlighting similarity/differences, various forms of probability and rule boundaries respectively (red arrows).

developers build more user-centric explainable AI-based systems with targeted XAI features.

We have previously introduced the conceptual framework in [42] and we refer interested readers to study the original work. In this position paper, we further demonstrate how to use the framework to design targeted explanation features, focusing on the choice of explanation types defined by Lim and Dey [24, 25]. Lim and Dey defined their explanation taxonomy for context-aware computing and described how users might use them [24] as well as empirically demonstrated the effectiveness of some explanation types [23]. However, in later studies, they found that users could also reason differently than anticipated and have different preferences for explanation types even for the same tasks [27, 29]. For example, to understand how a mobile app would behave in a new situation, Lim and Dey found that users could either use the What If explanation, or exercise the actual scenario and inspect the Why explanation.

Our conceptual framework in [42] provides theoretical support for why users preferred certain explanations in addition to also describing how people are subject to reasoning fallacies and cognitive biases as well as how to select explanations to mitigate these biases. In this work, we focus on supporting three explanation goals by providing examples for specific XAI features that support them and relate them to Lim and Dey's intelligibility types. To provide some context before we apply the framework, we summarize relevant parts of the original work [42] in the next section.

2 XAI Framework of Reasoned Explanations

We performed a literature review and synthesized a conceptual framework from rationalizing logical connections. Rather than perform a comprehensive encyclopedic literature review of relevant concepts in XAI [1, 5, 10, 43], our goal was to create an operational framework with which developers of XAI interfaces and systems can use. We started with an existing literature review of different XAI techniques [1] and considered what reasoning or cognitive theories would justify the need for such methods. In addition, we considered how people reason and make decisions so as to design XAI that normalizes to their thought processes; this minimizes the learning curve of XAI facilities. Our literature search was inspired by work from Miller [31], but we limited our scope to philosophy, cognitive psychology, and AI. We iteratively refined our framework by 1) finding concepts in XAI, reasoning and psychology, 2) drawing connections between them to elucidate relationships, 3) finding gaps to justify why certain XAI techniques could be useful, and 4) searching for more concepts.

We have developed a conceptual framework that links concepts in human reasoning processes with explainable AI techniques. By considering two aspects of the human and the machine, we further divide the framework into four main modules. We focus on two modules here. First, we identify how people ideally reason and why we seek explanations (2.1). These articulate reasoning methods and explanation types that provide the foundation of what good decision aids should support. Second, we describe various AI modeling and XAI facilities, and contextualize how they have been developed to support certain reasoning methods (2.2). Figure 1 shows these key modules and pathways linking them to illustrate how some reasoning methods can be supported by XAI facilities.

2.1 How People should Reason and Explain

This section informs how XAI can support different explanation types by articulating how people understand events or observations through explanations. We drew these insights from the fields of philosophy, and cognitive psychology, specifically 1) different ways of knowing, 2) what structures contain knowledge, 3) how to reason logically and 4) why we seek explanations.

2.1.1 Explanation Goals. The needs for explanations are triggered by a deviation from expected behavior [31], such as a curious, inconsistent, discrepant or anomalous event.

Alternatively, users may also seek to monitor for an expected, important or costly event. Miller identified that the main reason why people want explanations is to facilitate learning by allowing the user to (i) filter to a small set of causes to simplify their observation, and to (ii) generalize these observations into a conceptual model where they can predict and control future phenomena [31]. The latter goal of prediction is also described as human-simulatability [30]. We orient our discussion of explanations with respect to these broad goals of finding causes and concept generalization.

From the AI research perspective, a recent review by Nunes and Jannach summarized several purposes for explanations [32]. Explanations are provided to support ***transparency***, where users can see some aspects of the inner state or functionality of the AI system. When AI is used as a decision aid, users would seek to use explanations to ***improve their decision making***. If the system behaved unexpectedly or erroneously, users would want explanations for ***scrutability*** and ***debugging*** to be able to identify the offending fault and take control to make corrections. Indeed, this goal is very important and has been well studied regarding user models [3, 16] and debugging intelligent agents [21]. Finally, explanations are often proposed to improve trust in the system and specifically ***moderate trust*** to an appropriate level [4, 6, 26].

2.1.2 Inquiry and Reasoning. With the various goals of explanations, the user would then seek to find causes or generalize their knowledge and reason about the information or explanations received. Pierce defined three kinds of inferences [34]: deduction, induction, and abduction. **Deductive reasoning** "top-down logic" is the process of reasoning from premises to a conclusion. **Inductive reasoning** "bottom-up logic" is the reverse process of reasoning from a single observation or instance to a probable explanation or generalization. **Abductive reasoning** is also the reverse of deductive reasoning and reasons from an observation to the most likely explanation. This is also known as "inference to the best explanation". It is more selective than inductive reasoning, since it prioritizes hypotheses.

Popper combined these reasoning forms into the **Hypothetico-Deductive model** as a description of the scientific method [2, 35]. The model describes the steps of inquiry as (1) observe and identify a new problem, (2) form a hypothesis as induction from observations, (3) deduce consequent predictions from the hypotheses, and (4) test (run experiments) or look for (or fail to find) further observations that falsify the hypotheses. It is commonly used and taught in medical reasoning [8, 9, 33]. A key aspect of the HD model is hypothesis generation where observation of the current state can help the user decide whether to test for relationships between potential causes and the outcome effect.

Finally, **analogical reasoning** is the process of reasoning from one instance to another. It is a weak form of inductive reasoning since only one instance is considered instead of many examples [41]. Nevertheless, it is often used in case base reasoning and in legal reasoning to explain based on precedence (same case) or analogy (similar case) [22].

2.1.3 Causal Attribution and Explanations. As users inquire for more information to understand an observation, they may seek different types of explanations. Miller identified *causal explanations* as a key type of explanation, but also distinguished them from *causal attribution*, and *non-causal* explanations [31].

Causal attribution refers to the articulation of internal or external factors that could be attributed to influence the outcome or observation [11]. Miller argues that this is not strictly a causal explanation, since it does not precisely identify key causes. Nevertheless, they provide broad information from which users can judge and identify potential causes. Combining attribution across time and sequence would lead to a ***causal chain***, which is sometimes considered a trace explanation or line of reasoning.

Causal explanation refers to an explanation that is focused on the *selected* causes relevant to interpreting the observation with respect to existing knowledge. This requires that the explanation be ***contrastive*** between a *fact* (what happened) and a *foil* (what is expected or plausible to happen). Users can ask *why not* to understand why a foil did not happen. The selected subset of causes thus provides a ***counterfactual*** explanation of what needs to change for the alternative outcome to happen. This helps people to identify causes, on the scientific basis that manipulating a cause will change the effect. This also provides a more usable explanation than causal attribution, because it presents fewer factors (reduces information overload) and can provide users with a greater perception of control, i.e., *how to* control the system. A similar method is to ask *what if* the factors were different, then what the effect would be. Since this asks about prospective future behavior, Hoffman and Klein calls this ***transfactual*** reasoning; conversely, counterfactual reasoning asks retrospectively [13, 14]. This articulation highlights the importance of contrastive (Why Not) and counterfactual (How To) explanations instead of simple trace or attribution explanations typically used for transparency.

2.1.4 Summary. We have identified different inquiry and explanation goals, rational methods for reasoning, causal and non-causal explanation types, and evaluation with decisions to describe a chain of reasoning that people make. We next describe various explanations and AI facilities and how they support reasoning.

2.2 How XAI Generates Explanations

Now we turn to how algorithms generate explanations, in searching for connections with human explanation facilities. We characterize AI and XAI techniques by how they (1) semantically support human reasoning specific methods of scientific inquiry, such as Bayesian probability, similarity modeling, and queries; and (2) how to represent explanations with visualization methods, data structures and atomic elements. Where relevant we link AI techniques back to concepts (green text) in rational reasoning. **Bold** text refers to key constructs in each module in the framework, and *italic* text refers to sub-constructs.

2.2.1 Bayesian Probability. Due to the stochastic nature of events, reasoning with probability and statistics is important in decision making. People use inductive reasoning to infer events and test hypotheses. Particularly influential is Bayes theorem that

describes how the probability of an event depends on prior knowledge of observed conditions. This covers specific concepts of prior and posterior probabilities, and likelihood. Understanding outcome probabilities can inform users about the expected utility.

Bayesian reasoning helps decision makers to reason by noting the prevalence of events. E.g., doctors should not quickly conclude that a rare disease is probable, and they would be interested to know how influential a factor or feature is to a decision outcome.

2.2.2 Similarity Modeling. As people learn general concepts, they seek to group similar objects and identify distinguishing features to differentiate between objects. Several classes of AI approaches have been developed, including modeling *similarity* with **distance-based** methods (e.g., case base reasoning, clustering models), **classification** into different kinds (e.g., supervised models, nearest neighbors), and **dimensionality reduction** to find latent relationships (e.g., collaborative filtering, principal components analysis, matrix factorization, and autoencoders). Many of these methods are data-driven to match candidate objects with previously seen data (training set), where characterization depends on the features engineered and the model which frames an assumed structure of the concepts. Explanations of these mechanism are driven by inductive and analogical reasoning to understand why certain objects are considered similar or different. Identifying causal attributions can then help users ascertain the potential causes for the matching and grouping. Note that while rules appear to be a distinct explanation type, we could consider them as descriptions of the boundary conditions between dissimilar groups.

2.2.3 Intelligibility Queries. Lim and Dey identified several queries (called intelligibility queries) that a user may ask of a smart system [24, 25]. Starting from a usability-centric perspective, the authors developed a suite of colloquial questions about the system state (Inputs, What Output, What Else Outputs, Certainty), and inference mechanism (Why, Why Not, What If, How To). While they initially found that Why and Why Not explanations were most effective in promoting system understanding and trust [23], they later found that users may exploit different strategies to check model behavior and thus use different intelligibility queries for the same interpretability goals [26, 29].

2.2.4 XAI Elements. We identify several building blocks that compose many XAI explanations. By identifying these elements, we can determine if an explanation strategy has covered information that could provide key or useful information to users. This reveals how some explanations are just reformulations of the same explanation types but with different representations, such that the information provided and interpretability may be similar. Currently, showing feature ***attribution*** or influence is very popular, but this only indicates which input feature of a model is important or whether it had positive or negative influence towards an outcome. Other important elements include the ***name*** and ***value*** of *input* or *outputs* (generally shown by default in explanations, but fundamental to *transparency*), and the ***clause*** to describe if the value of a feature is above or below a threshold (i.e., a rule).

3 Intelligibility Types

We employ the taxonomy of Lim and Dey [24, 28] due to its pragmatic usefulness to operationalize in applications and to leverage the Intelligibility Toolkit [25] that makes it convenient to implement a wide range of explanations. While it does not currently generate recent state-of-the-art explanations and models, the explanation data structures allow it to be extended to support feature attribution and rules explanations. We reapply the original definitions to more general applications of machine learning beyond context-aware systems and introduce new types. We also describe and situate the explanation types in context of underlying reasoning processes.

Inputs explanations inform users what input values from data instances or sensors that the application is reasoning for the current case. When a user asks a why question, she may naively be asking for the Inputs state. We also consider this to be the basic form of explanation to support transparency by showing the current measured input or internal state of the application.

What Output explanations inform users what is the current outcome, inference, or prediction and what possible output options the application can produce. For applications that can have different outcome values (multiclass or multilabel), we can also show **Outputs** explanations. This lets users know what it can do or what states it can be in (*e.g.*, activity recognized as one of three options: sitting, standing, walking). This helps users understand the extent of the application's capabilities.

Certainty explanations inform users how (un)certain the application is of the output value produced. They help the user determine how much to trust the output value and whether to consider an alternative outcome. While originally, Lim and Dey considered the confidence outcome of a predictive model, this can now include stochastic uncertainty from Bayesian modeling approaches, which is essentially the posterior probability. Furthermore, we have found that users may reason with prior and conditional probability, so three types of uncertainty should be supported: prior, conditional, posterior.

Why explanations inform users why the application derived its output value from the current (or previous) input values. This is typically represented as a set of triggered rules (rule trace) for rule-based systems or feature attributions (or weights of evidences) for why the inferred value was inferred over alternative values. Compared to Input explanations, Why explanations focus on highlighting a subset of key variables or clauses, though this does not specifically support counterfactual reasoning, especially for multi-class classification systems.

Why Not explanations inform users why an alternative outcome (foil) was *not* produced, with respect to the inferred outcome (fact), given the current input values. Why Not explanations provide a *pairwise comparison* between the inferred outcome and an alternative outcome. Similar to Why explanations that help users to focus on key inputs, Why Not explanations focus on salient inputs that matter for contrasting between the fact and foil. With the fewer features highlighted, this can support counterfactual reasoning, where the user learns *how to* change key input values to achieve the alternative outcome. Hence, such

Why Not explanations are essentially **How To** explanations. Note that we can also interpret a Why explanation as a contrast between the inferred outcome and all other alternative outcomes. However, also note that Why Not explanations generated as feature attribution or weights of evidence are not particularly useful for How To explanations. As with Lim and Dey, we note that Why Not explanations are important to support, since users would typically ask for explanations when something unexpected happens, i.e., they expect the foil to happen. This agrees with Miller that most users truly ask for contrastive explanations (Why Not) and that these should be explained with counterfactuals (How To) [31].

What If explanations allow users to anticipate or simulate what the application will do given a set of user-set input values. While this straightforward explanation type has received little attention in recent AI research, it is an intuitive technique to support human simulatability defined by Lipton [30] and supports transfactual reasoning defined by Hoffman and Klein [13].

When explanations *(new)* indicate under what circumstance or scenario, or with what instance case would a particular outcome happen. This can be used to explain for inferred or alternative outcomes. Unlike Why or Why Not explanations which focus on input feature attributions or values, this focuses on the instance entity as a whole. Thus, it is suitable for exemplar, prototype or case-based explanations. Unlike How To explanations, this does not describe counterfactuals of a subset o inputs to change a scenario to have a different outcome. Note that we use a different definition as originally defined by Lim [28], which referred to the timestamp of the inference event.

4 Selecting Intelligibility for Explanation Goals

We had previously summarized several goals or reasons why people ask for explanations. These are primarily to improve their understanding of the AI-driven application, or the situation, or to improve their current or future ability to act predictably and correctly. In this section, we describe how to support three explanation goals — filter causes, generalize and learn, and predict and control — with the Intelligibility explanation types. By relating the use of these explanations explicitly back to user goals, we identify pathways to justify the use of various explanations in explainable AI. While our text and Figure 1 already describe some pathways, we articulate these clearer in this section.

4.1 Find and Filter Causes

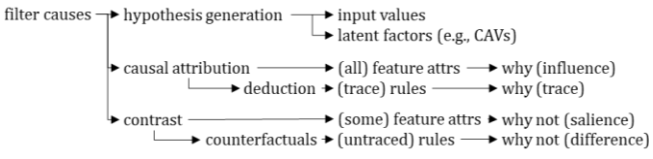

Figure 2. Pathways for using Inputs, Why and Why Not explanations to help the user to find and filter causes for the current system inference behavior.

We identified three pathways to help users narrow down and identify specific causes for a particular system outcome (see Figure 2). While Input explanations are most basic and colloquially queried by users, we identify that users would inspect the input feature values to find anomalies, discrepancies, or surprising values, then generate hypotheses for what could be wrong. This is not particularly efficient, since users are not directed to any salient cause, but it can allow users to determine their own hypotheses for causes. Going even further and giving users more choice for hypothesis generation, we can support the discovery of latent factors. While not originally defined in the Intelligibility framework [24, 25, 28], recent work by Kim et al. on TCAV [18] allows users to specify their own concepts of interest and test if they are influential in a model's inference.

The second pathway involves showing Why explanations as either a rule trace or feature attribution (or importance). This is driven by the users identifying the influence or attribution due to various causes (features) or by tracing deductive paths in the system rule logic.

The third pathway involves contrasting the inferred outcome (fact) with the expected outcome (foil). Salient large feature attribution differences can call the user's attention to potential causal features, but rules provide a more actionable method that explains how specific feature values could have led to the counterfactual case outcome.

4.2 Generalize and Learn

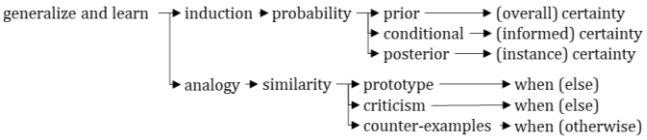

Figure 3. Pathways for using Certainty and When explanations to support users to generalize and learn about how the system would behave for similar cases.

We identified two main pathways to help the user learn a general mental model of how the system would behave (see Figure 3). The first pathway involves reasoning by induction, the user could be interested to know the likelihood of the outcome i) in general (overall), ii) the system's confidence or certainty of the outcome prediction for the current instance, or iii) an intermediate certainty where only some features matter (e.g., disease risk for all males, given that a patient is male).

The second pathway involves simpler but narrower reasoning by analogy, where the user looks at one instance at a time to form a detailed understanding of similar specific cases. Here, the system proposes the examples, such as i) prototypes to indicate median instances for each outcome type, ii) critique examples to indicate examples of a desired outcome that are close to the decision boundary [17], or iii) counter-examples that are similar to the current instance but have different predicted outcomes [38].

4.3 Predict and Control

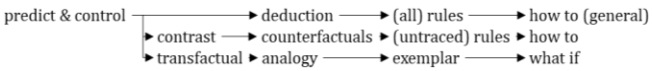

Figure 4. Pathways for using How To and What If explanations to support users to predict how the system would behave in a future case and to control its behavior for current or similar cases.

We identified three pathways to help users to predict the system's future behavior and control current behavior (see Figure 4). First, using deductive reasoning, users can read the full rule-set of the general How To explanation to predict how the system would inference. However, this can be tedious for a large rule-set. Second, focusing on a specific contrast case, users can use the How To counterfactual explanations of rule-based Why Not explanations to understand how they could try to change the situation for a different outcome. Anchors by Ribeiro et al. provide a good recent method for counterfactual explanations to support How To explanations [37]. Third, users could use a What If explanation to test specific instances that they are interested in; i.e., they set input values and observe the simulated outcomes. This is similar to When explanations, but the user chooses the input states and example.

Note that we consider explanations that build tree explainer models to be equivalent to rule-based explanations, since we can use first-order logic to transform them [25, 28]. Furthermore, we do not know of any feature attribution-based explanations that can specifically satisfy the explanation goal of prediction and control. The popularity of feature attributions presents a big gap in the research in XAI which tend to not produce actionable explanations.

5 Conclusion

We have described a theory-driven conceptual framework for designing explainable facilities by drawing from philosophy, cognitive psychology and artificial intelligence (AI) to develop user-centric explainable AI (XAI). Using this framework, we can identify *specific pathways* for how some explanations can be useful, how certain reasoning methods fail due to cognitive biases, and how to apply different elements of XAI to mitigate these failures. By articulating a detailed design space of technical features of XAI and connecting them with user requirements of human reasoning, our framework aims to helps developers build more user-centric explainable AI-based systems.

REFERENCES

[1] Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., Kankanhalli, M. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '18.*

[2] Anderson, H. (2015). Scientific Method. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/scientific-method/. Retrieved 10 September 2018.

[3] Assad, M., Carmichael, D. J., Kay, J., & Kummerfeld, B. (2007, May). PersonisAD: Distributed, active, scrutable model framework for context-aware services. In *International Conference on Pervasive Computing* (pp. 55-72). Springer, Berlin, Heidelberg.

[4] Antifakos, S., Schwaninger, A., & Schiele, B. (2004, September). Evaluating the effects of displaying uncertainty in context-aware applications. In *International Conference on Ubiquitous Computing* (pp. 54-69). Springer, Berlin, Heidelberg.

[5] Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)* (p. 8).

[6] Bussone, A., Stumpf, S., & O'Sullivan, D. (2015, October). The role of explanations on trust and reliance in clinical decision support systems. In *Healthcare Informatics (ICHI), 2015 International Conference on* (pp. 160-169). IEEE.

[7] Coppers, S., Van den Bergh, J., Luyten, K., Coninx, K., van der Lek-Ciudin, I., Vanallemeersch, T., & Vandeghinste, V. (2018, April). Intellingo: An Intelligible Translation Environment. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 524). ACM.

[8] Croskerry, P. (2009a). A universal model of diagnostic reasoning. *Academic medicine*, *84*(8), 1022-1028.

[9] Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1978). *Medical Problem Solving: An Analysis of Clinical Reasoning*. Cambridge, MA: Harvard University Press.

[10] Graesser, A.C., Person, N., Huber, J. (1992). Mechanisms that generate questions. In: Lauer, T.W., Peacock, E., Graesser, A.C. (Eds.), *Questions and Information Systems.* Lawrence Erlbaum, Hillsdale, NJ, pp. 167–187.

[11] Heider, F. (2013). The psychology of interpersonal relations. Psychology Press.

[12] Herlocker, J., Konstan, J., and Riedl, J. (2000). Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work (CSCW'00).* ACM, New York, NY, USA, 241-250.

[13] Hoffman, R. R., & Klein, G. (2017a). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, (3), 68-73.

[14] Hoffman, R. R., Mueller, S. T., & Klein, G. (2017b). Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intelligent Systems*, *32*(4), 78-86.

[15] Hoffman, R., Miller, T., Mueller, S. T., Klein, G., & Clancey, W. J. (2018). Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intelligent Systems*, *33*(3), 87-95.

[16] Kay, J. (2001). Learner control. *User modeling and user-adapted interaction*, *11*(1-2), 111-127.

[17] Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems* (pp. 2280-2288).

[18] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018, July). Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *International Conference on Machine Learning* (pp. 2673-2682).

[19] Klein, G. (2018). Explaining Explanation, Part 3: The Causal Landscape. *IEEE Intelligent Systems*, *33*(2), 83-88.

[20] Krause, J., Perer, A., & Ng, K. (2016, May). Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5686-5697). ACM.

[21] Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015, March). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 126-137). ACM.

[22] Lamond, G. (2006). Precedent and analogy in legal reasoning. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/legal-reas-prec/. Retrieved 10 September 2018.

[23] Lim, B. Y., Dey, A. K., & Avrahami, D. (2009a, April). Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2119-2128). ACM.

[24] Lim, B. Y., & Dey, A. K. (2009b, September). Assessing demand for intelligibility in context-aware applications. In *Proceedings of the 11th international conference on Ubiquitous computing* (pp. 195-204). ACM.

[25] Lim, B. Y., & Dey, A. K. (2010, September). Toolkit to support intelligibility in context-aware applications. In *Proceedings of the 12th ACM international conference on Ubiquitous computing* (pp. 13-22). ACM.

[26] Lim, B. Y., & Dey, A. K. (2011a, September). Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th international conference on Ubiquitous computing* (pp. 415-424). ACM.

[27] Lim, B. Y., & Dey, A. K. (2011b, August). Design of an intelligible mobile context-aware application. In *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 157-166). ACM.

[28] Lim, B. Y. (2012). Improving understanding and trust with intelligibility in context-aware applications. *PhD dissertation*. CMU.

[29] Lim, B. Y., & Dey, A. K. (2013, July). Evaluating Intelligibility Usage and Usefulness in a Context-Aware Application. In *International Conference on Human-Computer Interaction* (pp. 92-101). Springer, Berlin, Heidelberg.

[30] Lipton, Z. C. (2016). The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*.

[31] Miller, T. (2017). Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269*.

[32] Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3-5), 393-444.

[33] Patel, V. L., Arocha, J. F., & Zhang, J. (2005). Thinking and reasoning in medicine. *The Cambridge handbook of thinking and reasoning*, *14*, 727-750.

[34] Peirce, C. S. (1903). Harvard lectures on pragmatism, Collected Papers v. 5.

[35] Popper, Karl (2002), Conjectures and Refutations: The Growth of Scientific Knowledge, London, UK: Routledge.

[36] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144). ACM.

[37] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018a). Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*.

[38] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018b). Semantically Equivalent Adversarial Rules for Debugging NLP Models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (Vol. 1, pp. 856-865).

[39] Roth-Berghofer, T. R. (2004, August). Explanations and case-based reasoning: Foundational issues. In *European Conference on Case-Based Reasoning* (pp. 389-403). Springer, Berlin, Heidelberg.

[40] Silveira, M.S., de Souza, C.S., and Barbosa, S.D.J. (2001). Semiotic engineering contributions for designing online help systems. In *Proceedings of the 19th annual international conference on Computer documentation (SIGDOC '01)*. ACM, New York, NY, USA, 31-38.

[41] Vickers, John (2009). The Problem of Induction. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/induction-problem/. Retrieved 10 September 2018.

[42] Wang, D., Yang, Q., Abdul, A., Lim, B.Y. 2019. Designing Theory-Driven User-Centric Explainable AI. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '19. https://doi.org/10.1145/3290605.3300831

[43] Zhang, Q. S., & Zhu, S. C. (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 27-39.