Designing Theory-Driven User-Centric Explainable AI

Danding Wang[1], Qian Yang[2], Ashraf Abdul[1], Brian Y. Lim[1]
[1]School of Computing, National University of Singapore, Singapore
[2]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, United States
wangdanding@u.nus.edu, yangqian@cmu.edu, ashrafabdul@u.nus.edu, brianlim@comp.nus.edu.sg

ABSTRACT

From healthcare to criminal justice, artificial intelligence (AI) is increasingly supporting high-consequence human decisions. This has spurred the field of explainable AI (XAI). This paper seeks to strengthen empirical application-specific investigations of XAI by exploring theoretical underpinnings of human decision making, drawing from the fields of philosophy and psychology. In this paper, we propose a conceptual framework for building human-centered, decision-theory-driven XAI based on an extensive review across these fields. Drawing on this framework, we identify pathways along which human cognitive patterns drives needs for building XAI and how XAI can mitigate common cognitive biases. We then put this framework into practice by designing and implementing an explainable clinical diagnostic tool for intensive care phenotyping and conducting a co-design exercise with clinicians. Thereafter, we draw insights into how this framework bridges algorithm-generated explanations and human decision-making theories. Finally, we discuss implications for XAI design and development.

CCS CONCEPTS

• Human-centered computing ~ Human computer interaction

KEYWORDS

Intelligibility, Explanations, Explainable artificial intelligence, Clinical decision making, Decision making

ACM Reference format:
Danding Wang, Qian Yang, Ashraf Abdul, Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *2019 CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland, UK.* ACM, New York, NY, USA. **13** pages. https://doi.org/10.1145/3290605.3300831

CHI 2019, May 4-9, 2019, Glasgow, Scotland, UK.

ACM ISBN 978-1-4503-5970-2/19/05...$15.00.
DOI: https://doi.org/10.1145/3290605.3300831

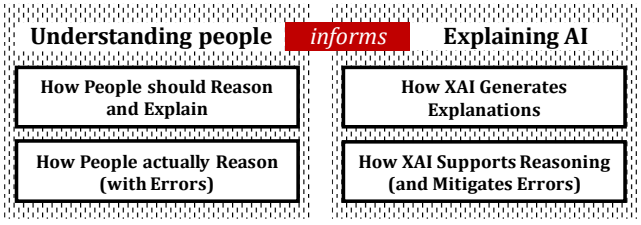

Figure 1. Overview of user-centric XAI framework.

1 INTRODUCTION

From supporting healthcare intervention decisions to informing criminal justice, artificial intelligence (AI) is now increasingly entering the mainstream and supporting high-consequence human decisions. However, the effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users in these critical situations. These challenges have spurred research interest in explainable AI (XAI) [2, 12, 32, 43, 109]. To enable end users to understand, trust, and effectively manage their intelligent partners, HCI and AI researchers have produced many user-centered, innovative algorithm visualizations, interfaces and toolkits (e.g., [18, 56, 67, 86] that support users with various levels of AI literacy in diverse subject domains, from the bank customer who is refused a loan, the doctor making a diagnosis with a decision aid, to the patient who learns that he may have skin cancer from a smartphone photograph of his mole [30].

Adding on to this line of inquiry, this paper seeks to strengthen empirical application-specific investigations of XAI by exploring theoretical underpinnings of human decision making, drawing from the fields of philosophy and psychology. We first conducted an extensive literature review in cognitive psychology, philosophy and decision-making theories that describe patterns of how people reason, make decisions and seek explanations, and cognitive factors that bias or compromise decision-making.

We drew connections between these insights and explanation facilities that AI algorithms commonly produce, and in turn proposed a theory-driven, user-centric XAI framework (Figure 1). With this framework, XAI researchers and designers can identify pathways along which human cognitive patterns drive needs for building XAI and how XAI can mitigate common cognitive biases. Next, to evaluate this framework by putting it to work, we applied it to a real-world clinical machine learning (ML) use case, i.e., an explainable diagnostic tool for intensive care phenotyping. Co-designing with 14 clinicians, we developed five explanation strategies to mitigate decision biases and moderate trust. We implemented the system with XGBoost [17] trained on the MIMIC III data [45]. Drawing on this application, we reflect on the utility and limitations of the framework and share lessons learned. **Our contributions are**:

1. A theory-driven conceptual framework linking different XAI explanation facilities to user reasoning goals which provides pathways to mitigate reasoning failures due to cognitive biases.
2. An application of our framework to medical decision making to demonstrate its usefulness in designing user-centric XAI.
3. Discussion to generalize our framework to other applications.

The key takeaway of the framework is to choose explanations backed by reasoning theories and cognitive biases rather than based on taxonomy (e.g., [32, 38, 66, 90, 93]) or popular XAI techniques (e.g. [75, 86]). This aims to help developers build human-centric explainable AI-based systems with targeted XAI features.

2 BACKGROUND AND RELATED WORK

Research on explainable AI has recently been burgeoning with many algorithmic approaches being developed in AI and many empirical studies on the impact of explanations in HCI. We refer the interested reader to several literature reviews [2, 12, 32, 43, 109]. We identified three approaches from prior literature and describe how our framework extends the explanation research landscape

1) Unvalidated guidelines for design and evaluation based on authors' respective experiences with little further justification [25, 59, 72]. The growing field of eXplainable AI (XAI) has produced algorithms to generate explanations as short rules [60, 63, 87], attribution or influence scores [15, 24, 74, 86], prototype examples, partial dependence plots [16, 56], etc. However, little justification is provided for choosing different explanation types or representations. It is unclear why these explanations will be feasibly useful to actual users. To better direct research efforts, some AI researchers have argued for rigorous goals and definitions of human interpretability, criteria of explanations and evaluation methods [25, 72]. Desirable properties include model faithfulness [59], simulatability [72], and empirical evaluation with human subjects [25]. However, it is unclear why these properties are important to users. These guidelines provide high-level objectives for designing explainable solutions, while our framework provides lower-level building blocks to construct these solutions.

2) Empirically derived taxonomies of explanation needs elicited from user surveys [66]. As AI researchers respond to the call for more human subjects evaluations [25] (e.g., [50, 80, 88]), this adds to the large body of work in HCI, Recommender Systems, and Intelligent User Interfaces that have long evaluated explainable interfaces with empirical studies (e.g., [15, 27, 37, 59, 65, 66, 69, 70, 102]). Some recent studies focused on evaluating what and how much to explain [18, 30]. Findings include only showing explanations that extend prior knowledge [18], and not to be too "creepy" by disclosing too much information [30]. Some researchers have taken an open-ended approach to probe users about their explanation needs under various contexts [27, 53, 66, 101]. To make sense of the variety of explanations, several explanation frameworks have been proposed for knowledge-based systems [32], recommender systems [38], case-based reasoning [90], intelligent decision aids [93], tutoring systems [32], intelligible context-aware systems [66], etc. These frameworks were mostly taxonomic and derived from diverse but limited user studies. Recently, there are efforts to formalize the process to elicit mental models and explanation needs (e.g., with a participatory design framework [27]). Subsequent to such elicitation, designers can use our framework to identify relevant reasoning goals and select corresponding explanation techniques. Furthermore, we identify theories in human thinking that drives the needs for different explanations thus providing theoretical justification.

3) Psychological constructs from formal theories to guide explanation facilities via literature review [40, 41, 42, 79]. Vermeulen et al. drew from theories in HCI and psychology on affordances to more precisely define feedforward interactions and inform their proper design in interactive systems [102]. Particularly relevant to our work are recent writings by Miller, Hoffman and Klein which discussed relevant theories from philosophy, cognitive psychology, social science, and AI to inform the design of XAI [40, 41, 42, 52, 79]. 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. [40, 41, 42] and Klein [52] 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 to operationalize this rich body of work in the context of XAI-based decision support systems for specific user reasoning goals. Our framework builds on these constructs to explicitly link human reasoning and AI domains to provide operational pathways to select appropriate explanations.

3 METHODOLOGY

We sought to understand how people reason and which XAI facilities could satisfy their reasoning goals. Given the mature literature on human reasoning and large volume of work on XAI techniques, we first applied a *rationalistic* epistemological [33, 77] approach as our method of inquiry to develop our framework. That is, we performed a literature review and synthesized a conceptual framework from rationalizing logical connections. We then show a real-world example of applying the resulting framework in a concrete application by implementing specific explanations and evaluating their use in a co-design exercise with real users. This applies an *empirical* epistemological approach to inquiry.

Literature review for forming the framework. Rather than perform a comprehensive encyclopedic literature review of relevant concepts in XAI [2, 12, 32, 109], our goal was to create an operational framework which developers of XAI interfaces and systems can use. We started with an existing literature review of different XAI techniques [2]. Specifically, we focused on medical diagnosis literature on medical decision making, philosophy and cognitive psychology. 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.

Application of framework by developing an explainable AI decision aid. The validity and usefulness of a theoretical framework can be difficult to evaluate. Most frameworks are validated through debates within the research community or simply by the test of time. As a first step to evaluate our proposed framework, we apply it to develop an explainable decision support tool for medical diagnosis. Medicine is a critical application domain with a long history of leveraging on AI and requiring explainability [3, 10, 92]. We evaluated the designed explanations with users (clinicians) and implemented a real ML model with a real clinical dataset, to avoid spurious reasoning and insights that may arise from overly simplified or unrealistic medical scenarios.

4 XAI FRAMEWORK OF REASONED EXPLANATIONS

We have developed a conceptual framework (Figure 2) 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. First, we identify how people ideally reason and why we seek explanations (Section 4.1). 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 (4.2). Third, we introduce caveats in human decision making, highlighting how cognitive biases impair our reasoning (4.3). Fourth, we describe how XAI can be leveraged to mitigate decision biases (4.4). Figure 2 shows the key modules and pathways linking them to illustrate how some reasoning methods can be supported by XAI facilities.

4.1 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, psychology, and decision science, specifically 1) different ways of knowing, 2) what structures contain knowledge, 3) how to reason logically and 4) why we seek explanations.

4.1.1 Explanation Goals. The needs for explanations are triggered by a deviation from expected behavior [39, 79], such as a curious, inconsistent, discrepant or an anomalous event. Alternatively, users may also seek to monitor for an expected, important or costly event. Miller [79] summarized the main reason that people want explanations is to facilitate learning by allowing the user to (i) filter to a small set of causes to simplify their observation [73], and to (ii) generalize these observations into a conceptual model where they can predict and control future phenomena [37]. The latter goal of prediction is also described as human-simulatability [72]. We orient our discussion of explanations with respect to these broad goals of finding causes and concept generalization.

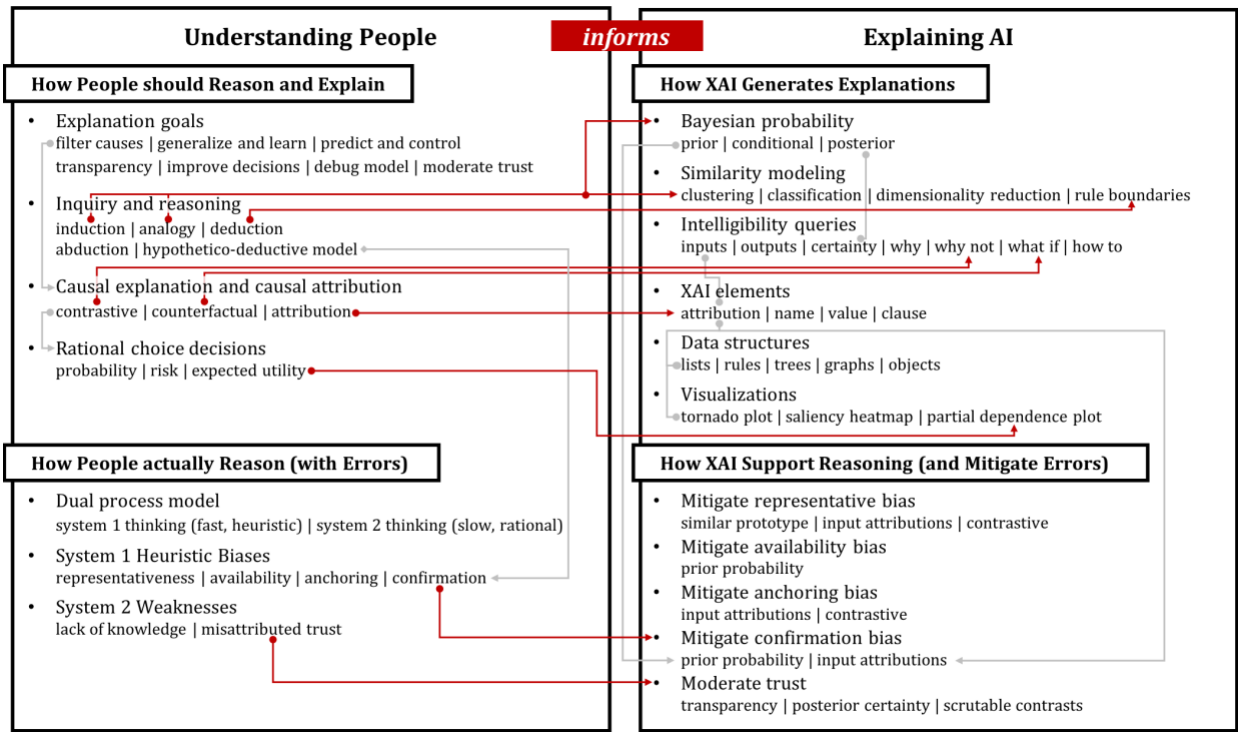

Figure 2. Conceptual framework for Reasoned Explanations that describes how human reasoning processes (left) informs XAI techniques (right). Points describe different theories of reasoning, XAI techniques, and strategies for designing XAI. Arrows indicate pathway connections: red arrows for how theories 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, hypothetico-deductive reasoning can be interfered by System 1 thinking and cause confirmation bias (grey arrow). Confirmation bias can be mitigated (follow the red line) by presenting information about the prior probability or input attributions. Next, we can see that input attributions can be implemented as lists and visualized using tornado plots (follow the grey line). We identified many pathways but show only a few for brevity.

From the AI research perspective, a recent review by Nunes and Jannach summarized several purposes for explanations [81]. 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 important and has been well studied regarding user models [7, 48] and debugging intelligent agents [59]. Finally, explanations are often proposed to improve trust in the system and specifically ***moderate trust*** to an appropriate level [8, 15, 68].

4.1.2 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 [83]: 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 [5, 84, 107]. 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 [20, 28, 82].

Finally, **analogical reasoning** as 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 [95, 103]. Nevertheless, it is often used in case base reasoning and in legal reasoning to explain based on precedence (same case) or analogy (similar case) [61, 96].

4.1.3 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 [79].

Causal attribution refers to the articulation of internal or external factors that could be attributed to influence the outcome or observation [37]. 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 considered a trace explanation or line of reasoning.

Causal explanation refers to an explanation that is focused on *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) [71, 79]. 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 principle 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 [40, 41]. 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.

4.1.4 Decision Theories. Several theories from decision science describe how to make decisions to optimize choice. Rational choice theory, and specifically, expected utility [105, 106], describes how people make decisions under uncertainty based on the utility (value or risk) and probability of different outcomes. By understanding the relative value of outcomes, people can compare outcomes and determine causes for why some outcomes were valued higher than others, or why their probabilities differed.

4.1.5 Summary. We have identified theoretical bases of inquiry, reasoning and explanation goals. We next describe various explanations and AI facilities and how they support reasoning.

4.2 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 (A) support human reasoning and specific methods of scientific inquiry, such as Bayesian probability, similarity modeling, and queries; and (B) how to represent explanations with visualization methods, data structures and atomic elements. Where relevant we link AI techniques back to concepts (blue text) in rational reasoning.

4.2.1 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, specifically, prior and posterior probabilities, and likelihood [9]. 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.

4.2.2 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 [1], clustering models [76]), **classification** into different kinds (e.g., supervised models [e.g., 19, 85], nearest neighbors [4]), and **dimensionality reduction** to find latent relationships (e.g., collaborative filtering [13], principal components analysis [44], matrix factorization [55], and autoencoders [91, 104]). 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 mechanisms 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 consider them descriptions of the boundary conditions between dissimilar groups.

4.2.3 Intelligibility Queries. Lim and Dey identified several queries (called intelligibility queries) that a user may ask of a smart system [66, 67]. 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 [65], they later found that users use different strategies to check model behavior and thus use different intelligibility queries for the same interpretability goals [68, 70]. In this work, we identify theoretical foundations that justify these different queries.

The previous subsections described mechanisms and constructs driven by reasoning semantics and explanation goals. These generate explanations that are often in similar forms. Next, we describe common data structures and atomic elements of explanations that are used to represent these semantic structures.

4.2.4 XAI Elements. We identify 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 popular is showing feature ***attribution*** or influence [16, 56, 74, 86] by indicating which input feature of a model is important or whether it had positive or negative influence towards an outcome. Another intuitive approach highlights similar (or different) **instances** from the training data, prototypes, criticism, and counter-examples [49, 54, 88]. Other 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).

4.2.5 (Data) Structures. With the elements identified, we next describe several structures commonly used. These generally align with concepts in data structures that are taught in undergraduate computer science. The simplest and most common way to construct explanations is as ***lists*** [16, 56, 74, 86]. If sorted, this would represent concepts related by some criteria of importance. Lists are often used to represent input feature attributions, and can also represent output class attributions. Logical clauses can be combined into ***rules*** [63] or as a ***decision tree*** to describe branching logic. To represent more complicated relationships or conceptual models, such as ontologies, ***graphs*** with nodes and edges are used. In non-CS disciplines, such as education, these are sometimes called *concept maps*. Essentially, these structures describe concepts and their relationships. Executing these rules would exercise deductive reasoning, a key method of inquiry. Finally, the extensible ***object*** data type can be used to represent simple (e.g., linear equation) or arbitrary structures that are domain dependent. This is useful for representing *exemplar* or *prototype* instances of concept classes or cases in case-based reasoning. By abstracting concepts into higher level generalizations, using objects in explanations leverages on analogical and inductive reasoning.

4.2.6 Visualizations. While some explanation structures can be communicated with textual explanations or single lists, complex concepts are better explained with visualization techniques. To provide data and algorithmic transparency, basic charts can be used to represent raw data (e.g., ***line charts*** for time series data) and canonical visualizations can be used to illustrate model structure (e.g., ***node-link diagrams*** for trees and graphs). To support causal attribution, ***tornado diagrams*** of vertical bar charts are popularly used for lists of attributions [86], ***saliency heatmaps*** for attributing to pixels or super-pixels for image-based models [50, 86], or highlighting on paragraph text [15, 58]. These visualization techniques support contrastive explanations and counterfactual reasoning by allowing for the *comparison* of different attributions (e.g., bar charts, heatmaps) or understanding of *relationships* between factors (tree or graph visualizations). Leveraging on inductive reasoning, ***scatterplot diagrams*** help users to perceive the similarity between objects by presenting objects in lower dimensionality projection (e.g., t-SNE) [47]. Drawing from statistical inference, ***partial dependence (PD) plots*** have been used to visualize how feature attribution varies across different feature values [16, 56]. Some work has extended this to show interaction effects between two factors [16, 75]. Finally, ***sensitivity analysis*** provides several methods as a robustness test to ask *what if* the input factors change slightly or are perturbed, whether the outcome of a decision will change. These visualization techniques support *rational choice* reasoning with counterfactuals by displaying how the

expected utility or risk of an outcome changes as input factors change. Hence, these decision aids, if used properly and deliberately, can support rational decision making.

4.2.7 Summary. We found that different AI models adhere to different human reasoning processes, and many recent XAI explanations are composed of common building blocks that can be represented by interpretable data structures and visualizations. These relationships can guide the choice of XAI techniques and components to help compose coherent explanation solutions depending on the explanation goals and expected reasoning methods users.

4.3 How People Actually Reason (with Errors)

So far, we have described how people reason rationally and the XAI explanations that can support this. However, people also reason heuristically to make decisions faster. In this section, we provide an overview of how people reason quickly and how decision errors can arise. This is best described with the *dual process model* [46], which has been used to universally describe many aspects of cognitive psychology. To make our analysis more concrete, we draw from literature in medical decision making.

4.3.1 Dual Process Model. Croskerry described a universal model for diagnostic reasoning as primarily driven by the Dual Process Model, which was popularized by Kahneman [20, 21, 22]. This clusters different approaches of thinking into System 1 and System 2, thinking fast and slow, respectively [46].

System 1 thinking is fast, intuitive, and low effort, where the person employs heuristics (unconscious cognitive shortcuts) to make quick decisions. For example, to quickly recognize an item, we apply the *representativeness* heuristic to use inductive reasoning to compare it and determine its similarity to a previously observed item. Thus, an experienced person who has seen many examples and learned to generalize can make decisions quickly with this pattern matching approach. However, as we discuss later, this can also lead to cognitive bias for inexperienced people.

System 2 thinking is slow, analytical, and high effort, where the person employs rational reasoning processes. Given the structure required for System 2 thinking, people need to be trained in their specific domain to have the necessary logical, semantic or mathematical concepts to reason. We had described processes and models from philosophy and AI in Sections 4.2 and 4.3, such as deductive reasoning, the hypothetico-deductive model, and Bayesian reasoning. These are relevant methods that people can systematically use under System 2 thinking. However, it is well-known that System 1 processes can interfere with System 2 thinking such that instead of reasoning with rational choice, users may reason under ***bounded rationality*** [93], be subjected to ***prospect theory*** of risk aversion [100], etc.

These confounded reasoning processes are well documented [11]. In fact, even as the dual process model explains how people make good decisions quickly or rationally, it also explains why people make suboptimal or wrong decisions. We next describe some ways that decision errors arise.

4.3.2 System 1 Decision Errors: Heuristic Biases. Although heuristics can speed up decision making, they sometimes oversimplify the situation and are compromised by factors, such as overconfidence, fatigue and time pressure. Lighthall et al. described decision making in intensive care and highlighted several heuristic biases [64]. For brevity, we summarize them in Table 1.

4.3.3 Forward and Backward Reasoning with Hypothetico-Deductive Model. While System 2 thinking can lead to good decisions, System 1 processes can interfere with System 2 processes and lead to wrong decisions [20]. Patel and colleagues identified two types of reasoning due to differences in the levels of expertise of doctors [6, 82]. Expert doctors tended to use a forward-oriented, data-driven reasoning strategy, where they start with observation findings (vital signs, lab test results, etc.) and derive a hypothesis (diagnosis). In contrast, novice doctors tend to use a backward-oriented, hypothesis-driven reasoning strategy, where they first generate hypothesized diagnoses and then look for findings that match the hypothesis. Backward reasoning tends to lead to diagnostic errors, due to a lack of a knowledge base in novice doctors [82], and *confirmation bias*, where the decision maker tends to look for confirming evidence of a wrongly-chosen hypothesis [23]. Even with a rigorous reasoning model based on a reasonable knowledge base, this heuristic bias can lead to poor decisions. Later, we discuss how explanations can mitigate this problem and how poorly designed XAI may even aggravate it.

4.3.4 System 2 Decision Errors: Misattributed Trust in Wrongly Calibrated Tool. Even without cognitive biases, System 2 thinking may still fail due to using a miscalibrated tool [21] or a lack of domain knowledge [82]. Croskerry describes how a doctor depending unknowingly on a miscalibrated diagnostic test can lead to misdiagnosis and that the doctor should not have been confident in the tool [21]. In the context of AI-driven decision support, this is equivalent to decision makers depending on a model with

poor accuracy. Therefore, they over-trust an unreliable model.

4.3.5 Summary. We have described how cognitive heuristics can speed up and reduce effort in decision making but can also lead to non-rational reasoning or biased decisions. Although XAI techniques have implicitly been designed to support rational reasoning, in the next section, we discuss how to use our framework to design strategies to mitigate some cognitive biases.

4.4 How XAI Supports Reasoning (and Mitigates Errors)

Decision aids can be used to improve decision making and reduce decision errors. Here, we discuss how specific XAI facilities can be used to mitigate the aforementioned cognitive biases. Once again, we draw from literature in medical decision making to make our analysis more concrete. In addition to describing important heuristic biases, Lighthall et al. summarized strategies to overcome them [64] (Table 1). We selected a subset of heuristic biases for which we identify how XAI can play a role to mitigate them, and hence improve decision making and trust. Note that we do not provide a comprehensive solution, since there are over 50 known cognitive heuristics [11]. While these mitigating strategies are not new, we note that they are typically framed as educational approaches to promote meta-cognition in healthcare practitioners and increase awareness of decision making pitfalls [20, 64]. Instead, we propose that explanation facilities can be used as a crutch to help decision makers validate their thinking. We link explanation strategies back to XAI facilities (orange text).

4.4.1 Mitigate Representativeness Bias: Prototype Cases of Decision Outcomes. Representativeness bias happens when a decision maker perceives the current situation as similar to other cases of a wrong classification. This can be due to a lack of experience in seeing many examples or a lack of focus on salient features. To mitigate this bias, we can show ***prototype*** instances to represent different outcomes; the prototypes can be rank-ordered by their *similarity* to the current case or by explicitly showing a (dis)similarity *distance* metric. To allow for comparison between cases by inspecting features, the *difference in value or attribution* per feature can be shown to ***contrast*** the differences between cases.

4.4.2 Mitigate Availability Bias: Prevalence of Decision Outcomes. Availability bias can occur when decision makers are unfamiliar with how often a particular outcome happens. This can be mitigated by showing the base rate of the outcome (***prior probability***) based on frequencies in the training dataset, SHAP bias [74]. Key insight: this is contrary to showing system confidence or uncertainty, which is related to the posterior probability of the outcome.

4.4.3 Mitigate Anchoring Bias: Premortem of Decision Outcome. The anchoring bias occurs when the decision maker forms a skewed perception due to an anchor and fixates on a decision (early closure). This limits the exploring of alternative hypotheses. To mitigate this, Lighthall et al. proposed highlighting how (inputs) findings may also be indicative of other hypotheses [64]. This can be facilitated by showing ***input attributions*** for multiple outcomes. By seeing how attributions are ***contrasted*** for different hypotheses, doctors may identify alternative diagnoses.

Lighthall et al. also proposed using Klein's premortem prospective hindsight exercise [51], where decision makers posit that their primary hypothesis ends up being wrong and have to determine why [51]. This can be facilitated by generating a ***counterfactual*** explanation with rules (e.g., Anchor LIME [87] or LORE [34]) to determine what input features could be slightly different to lead to a different outcomes. Alternatively, *sensitivity analysis* could be used to simulate a range of perturbed input values and test the stability of the primary hypothesis.

4.4.4 Mitigate Confirmation Bias: Discourage Backward-Driven Reasoning. To mitigate confirmation bias, we should encourage doctors to reason with the *hypothetico-deductive* method. We propose an explanation strategy that avoids backward-oriented reasoning by showing Findings (***input attributions***) first, instead of Hypotheses (***posterior probability*** or *class attribution*). Key insight: this is contrary to first showing system uncertainty as a shallow level of explanation and more transparent Inputs or mechanism as details on demand (e.g., [15, 69]). In addition, to encourage disproving their initial diagnosis, the system can show the ***prior probability*** of the inference (diagnoses).

4.4.5 Moderate Trust: Exposing System State and Confidence. Earlier work has argued for and demonstrated how providing explanations can improve trust because of increased user understanding (e.g., [65, 69, 81, 98]), however, Lim and Dey have also found that users should not trust systems when they are poorly performing; showing explanations of low confidence can further decrease this trust [69]. From our framework, we believe that this over-trust is due to biased System 2 thinking of fallaciously believing that the AI system has a higher-than-expected accuracy. To moderate users to trust the system at appropriate times or trust specific sub-components of the

Table 1. Heuristic biases that lead to decision (diagnostic) errors and strategies for mitigating them as reported in [64]. We consider how XAI can provide facilities to aid in these strategies.

Heuristic Bias	Description	Strategies to overcome systematic errors [64]	XAI Strategies for Medical Decisions
Represent-ativeness	Judging likelihood of an event 'A' belonging to a condition due to similarities between the two, but not judging whether A belongs to some other process that could be more similar.	**Compare** disease with **prototypes** of the condition; be suspicious when there is **no good match.**	- Identify **prototypes** of patient instances for each diagnosis - Show similarity between current patient and prototype(s) via **similarity distance**. - Highlight similarity and **contrast** differences in terms of data feature value or **attributions**.
Availability	Bias in perceiving that memorable, unusual or adverse events are more likely (frequent) than they truly are.	Seek **base rate** of a diagnosis.	- Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).
Anchoring	Skewed perception of a value due to a supplied numerical value (anchor).	Avoid confirmation and early closure; make use of **lab tests to "prove" other** leading diagnoses. "Crystal ball" exercise (**"premortem"** prospective hindsight [51]).	- Show input **attributions** for multiple outcomes to allow *contrastive reasoning.* - Facilitate **counterfactual** to test *How To* reduce the probability of primary diagnosis with Rules (e.g., aLIME, LORE). - Facilitate **sensitivity analysis** with *What If* explanations to test stability of primary hypothesis.
Confirmation	Collecting redundant information to confirm an existing hypothesis, instead of finding evidence of competing possibilities.	- Use **hypothetical-deductive** method to assess value and role of contemplated tests. - Try to **disprove your diagnosis**, consider conditions of **higher prevalence**.	- Show Findings (**input attribution**) first, instead of Hypotheses (**output posterior probability**). ***Insight:*** *this is opposite to typical Machine Learning apps to show output uncertainty first.* - Show **prior probability** (equivalent to SHAP bias) of diagnoses (in dataset).

system [70], the system should be ***transparent*** to show What and Inputs explanations and show its classification ***certainty***. Once the user detects erroneous reasoning in the system, *scrutability* features can be ***contrasted*** to allow for model debugging and correction (by a technical expert).

4.4.6 Summary. We have described how we can use various XAI components of our framework and identify pathways to support good reasoning behaviors in users. This seeks to mitigate cognitive biases that may arise due to heuristic reasoning.

4.5 Summary of Framework

Our framework describes how people reason rationally (Section 4.1) and heuristically but subject to cognitive biases (4.3), how XAI facilities do support specific rational reasoning processes (4.2), and can be designed to target decision errors (4.4). The framework identifies pathways between human reasoning and XAI facilities that can help organize explanations and identify gaps to develop new explanations given an unmet reasoning need. XAI application developers can use our framework as follows: First, consider the user's reasoning goals (4.1) and biases (4.3) for their respective apps. This can be informed through literature review, ethnography, participatory design [27], etc. Next, identify which explanations help reasoning goals (4.2) or reduce cognitive biases (4.4) using pathways in the framework (Figure 2 red arrows). Finally, integrate these XAI facilities to create explainable UIs. Furthermore, XAI researchers can extend our framework by 1) examining new XAI facilities (4.2) to understand how they were inspired by, dependent on, or built from reasoning theories (4.1) and 2) identifying common biases and reasoning errors (4.3) related to reasoning theories (4.1) and then identifying appropriate mitigation strategies (4.2) to select specific XAI facilities. For example, Informal Logic [99] could be integrated into reasoning theories (4.1) and informal fallacies [35] into reasoning errors (4.3).

5 APPLICATION: EXPLAINING MEDICAL DIAGNOSIS

We have developed our framework using a theory-driven approach and we next apply it by implementing an explainable AI-based early decision aid to diagnose patients in an Intensive Care Unit (ICU). Given the critical importance of ensuring correct medical diagnostic decisions, it is paramount for AI-based clinical decision support tools to explain themselves. Thus, the medical

domain is canonical for XAI (e.g., Leeds Abdominal Pain System [3], MYCIN [14, 92], DXPLAIN [10], GA^{2}M [16], Bayesian Rule List [63, 94], Prospector [56], EMBalance [15], AM-FM decompositions [43]). While these works use compelling explanations, the justification for selecting specific explanations is unclear. We demonstrate how using our framework, we can identify specific explanation types to improve medical decision making by reducing cognitive biases.

5.1 Implemented Model & Explanation Sketches

Using steps described in Section 4.5, we developed an explainable clinical decision support visualization to help reduce diagnostic error due to four cognitive biases (Table 1). We aim to explore how real users would interact with realistic explanations generated from a real model built over a real dataset to capture any nuance from AI decisions. Furthermore, given the complex nature of medical diagnosis, it is not reasonable for clinicians to reason over fake clinical cases as it could impair our findings if they encounter impossible details (e.g., vital signs not matching a diagnosis). Hence we used the MIMIC III [45] dataset to train a multi-label gradient boosted tree (XGBoost [17]) using 833 extracted features (17 patient vital × 7 time windows × 7 statistics) [36] to perform phenotype diagnosis based on the first 24 hours vitals of a patient in the ICU. We used state-of-the-art XAI facilities, such as SHAP [74] for attribution, LORE [34] for counterfactual rules, MOEA/D for sensitivity analysis [108] and implemented visualizations in Tableau. For simplicity, we presented most explanations aggregated in terms of 17 vital signs instead of the full 833 features. Figure 3 shows some examples of the explanation sketches which support different mitigation strategies. We iterated the visualizations by pilot testing with 5 clinical collaborators (not participants).

5.2 Co-Design Method and Participants

We recruited 14 clinicians (10 female; ages 24-29) from a local hospital: two senior residents, one medical officer and other residents. On average, they had 2.9 years of working experience, 2.6 years using a clinical decision support system and 2.1 months on rotation in the Intensive Care Unit (ICU). All participants were from the department of advanced internal medicine.

Each co-design session was done with one clinician and at least two research team members. After signing a consent form and agreeing to be audio recorded, we asked the participant to fill out a brief background survey. We briefed the participant on how to use and interpret the diagnosis dashboard and explanation facilities. We asked participants to diagnose one or two pre-selected patient cases using as many or little of the dashboard features as they wished. We understand that there may be erroneous explanations or model predictions, but this was beyond the scope of our initial exploration. We instructed participants to think aloud to explain their diagnostic reasoning process and their thoughts as they used the explanation interface. To investigate the usefulness of particular features, we occasionally prompted participants to consider different explanation facilities. To respect the busyness of doctors, each session was limited to about 30 minutes. Participants were compensated with café gift cards.

5.3 Findings from Co-Design

We consolidated findings from the co-design sessions and conducted an affinity diagram exercise with three research team members to cluster our notes into recurring themes. Our findings show how users' reasoning goals and meta-cognitive awareness of decision fallacies drive the (dis)use of different explanation types. With respect to the framework, the co-design exemplifies how to apply it by using its pathways to design an explainable interface and how to use its various constructs to interpret the reasoning processes that users engaged in.

5.3.1 Finding Alternative Hypotheses. We found many instances of users seeking alternative hypotheses with or without various types of XAI. First, most users deliberately did not want to be influenced or primed by the AI diagnosis prediction. E.g., P13 chose to start with seeing Feature values (vitals) and attributions, instead of seeing class prediction and attribution (inferred diagnosis and risk), because he *"would like to get an idea of [his] own differentials first instead of being influenced by system prediction"*. Users did not yet trust the AI's accuracy and were aware that this dependency may lead to decision bias. Indeed, this follows our understanding that backward reasoning can lead to confirmation bias. Users did not want to see any diagnostic prediction, risk score (class attribution), feature attribution, or any other explanations. Only one user, P3, demonstrated flawed **backward reasoning** and led his exploration by first inspecting class attributions and then feature attributions. He did this for his second patient case after seeing how the system was accurate with the first case. This led to fallacious diagnoses in the second case.

However, even though users initially avoided explanations, after they had made their initial hypothesis (diagnosis), to consider alternatives, they wanted to see what the AI decided. Users were mainly interested in (1) class attributions (predicted risk) or (2) feature attributions (vital

Figure 3. Screenshot of the AI-driven medical diagnosis tool with explanation sketches showing a patient with high predicted risk of acute myocardial infarction (AMI), heart disease, diabetes with complications, shock, etc. Explanations include (top left) feature value time series, (top right) class attribution of predicted disease risk, (middle right) feature attribution by vitals, and (bottom) counterfactual rules indicating key rules for each prediction. Interpretation: e.g., explanations suggest that the AI thinks that the patient has shock because of low oxygen saturation and blood pressure.

signs). If users saw a **class attribution** that was higher than expected (*contrast*), they would look for potentially responsible feature values and consider their validity. They would not necessarily see feature attributions, preferring to check feature values themselves. An alternative approach focused on **feature attributions** that were (2a) high or (2b) different than expected. In the former case, the high magnitude drew their attention to inspect the feature values, where they would generate other hypotheses. In the latter case, users had a mental model of what features should be important and they would *contrast* this with the attributions generated by the AI. This contrastive analysis behavior was also observed with **counterfactual rules**, where users were looking for unexpected rules.

We observed users applying analogical reasoning to see explanations of previous **prototypes**. P4, P7, and P14 wanted to see past patients which had similar presentations (e.g., complaints, vital signs), but not necessarily similar diagnoses (decision outcomes). Users were not interested in a quick summary or ranking of prototypes by an aggregate similarity score, since they do not trust the basis of the similarity. Instead, they would rather compare *feature values and attributions* to determine the extent of similarity. Users would then compare the ground truth or prediction outcomes of the past prototypes to consider alternatives.

As users sought out new alternative hypotheses, some also applied the *premortem* technique of discounting their primary diagnosis, i.e., assuming that it could be wrong and finding another viable alternative. P4 found counterfactual reasoning useful even though one cannot change a patient's history, *"because it's a matter of possibility. It's how you rank your differentials"*, indicating that she was open to alternative hypotheses. Users were interested to use *sensitivity analysis* to check on the stability of a diagnosis prediction by asking **counterfactual What If** questions and perturbing input values, seeing a **partial dependence** plot, or reading through the list of **counterfactual rules**. Given the long rules, some users found them tedious to read, while one user could rapidly skim them as she was mainly looking for unexpected rules

Overall, users iteratively performed forward and backward reasoning to generate and confirm hypotheses, following the *hypothetico-deductive model*. XAI facilities can provide several hints for further investigation, but most users prefer to return to the feature data or source material to confirm new hypotheses.

5.3.2 Coherent vs. Isolated Attribution. Users felt that the tornado plots suggested that each feature was independently attributing to a prediction. For example, P6 found it contradicting that the diastolic and systolic blood pressure features had opposing attributions for a heart attack. Users preferred to see related features together or how features may interact.

5.3.3 Lack of Trust: Verify with Supplementary, Situational, Source Data. A lack of trust drove users to want to verify the AI's decision by looking at raw data, in this case, patient case and physical exam notes, or recorded data, such as ECG readings, even if the AI model did not use such data. P7 felt that *"without the context of the patient, [she] can't diagnose based on the vitals itself. So [she] wouldn't really trust"*. In fact, this need to show raw situational data is commonly expected in different applications, such as showing test images in image recognition applications, and providing raw audio recordings in sound-based applications [70].

5.3.4 Conditional Prevalence of Decision Outcomes. To control availability bias, users were interested to see the prevalence of diagnosis. P7 used this as a hint to "bring [her] to rule the most prevalent cause out, but it won't be the determining factor whether [she] would like to

diagnose the patient". Specifically, they wanted to see prevalence for a subset of patients of most similar patients, including patients with the same demographics or presentation (vital signs, complaints, etc.). This demonstrates Bayesian reasoning, where doctors consider how knowing certain feature values can affect the posterior probability of the diagnosis.

6 DISCUSSION: LESSONS LEARNED FROM FRAMEWORK

Although our design of XAI strategies were theory-driven, our co-design exercise with real users provided insights into 1) how some XAI strategies were used as expected, and 2) how certain reasoning methods and goals were supported with unexpected different XAI facilities, and 3) how users were interested in additional XAI facilities or using them in ways that we did not previously consider. We discuss recommendations on how to further refine XAI designs to improve human interpretability. While our application was specific to the medical domain, we believe that these recommendations can apply to other domains.

Support hypothesis generation. A user seeking an explanation may know that the system output is not as expected and therefore seek a contrastive explanation [71, 79]. The explanation may suggest one or more causes, but the user will reason abductively to determine the most likely cause (best hypothesis). This process can involve repeated hypothesis generation and seeking of alternative hypotheses. We have found multiple XAI facilities that support hypothetico-deductive (H-D) reasoning, namely, feature and class attributions of the current instance, contrastive explanations of prototype instances, and counterfactual explanations with key rules and visualizations. While Miller argues for supporting contrastive explanations with counterfactual reasoning rules [71, 79], we further argue that XAI should support abductive and H-D reasoning i.e., in addition to providing counterfactuals to help users find causes, we should provide explanations to allow users to generate and test hypotheses to further narrow down potential causes.

Support forward (data-driven) reasoning by showing feature values and attributions before class attribution to avoid confirmation bias and backward reasoning, where the user does not consider other hypotheses independently. This recommendation runs counter to most recommendations of showing shallow explanations first and allowing details on demand, but mainly to limit information overload (e.g., [27, 58, 69]).

Support coherent factors. Depending on the domain, users may expect some features to be correlated or have some other relationship and would be confused if these features contradict their typical relationship. Such feature attributions should be aggregated together or have their interaction relationship visualized (e.g., with an interaction-effects partial dependence plot in GA^{2}M [16] or SHAP interaction plots [74]).

Supporting access to source and situational data. When adopting a new AI, users may want to manually perform decision making with a few instances to build up their trust. Showing raw data or supplementary data about the situation, even if not used directly by the model, can help with this verification goal. Example situational data include raw images and audio clips (e.g., [70]).

Support Bayesian reasoning. Many studies on explanations argue for showing system uncertainty [8, 68]. This is equivalent to showing the system posterior probability or class attribution. However, we found that it is also important to show other probabilities, namely, (1) prior probability to indicate the prevalence of classes in general, and (2) intermediate posterior probabilities, where after filtering on a set of salient features or factors to indicate the conditional prevalence of an outcome.

Integrating multiple explanations. Similar to [69, 70], we found that users employed a diverse range of XAI facilities, to reason variedly. Therefore, while much work has focused on developing good specific XAI algorithms, more work is needed to integrating multiple explanations into single explanations. Examples of multi-explanation applications include Laksa [70] and Intellingo [18].

7 LIMITATIONS AND FUTURE WORK

While our framework provides insights into using XAI facilities to target specific human reasoning, it does not cover all aspects of reasoning or all XAI techniques. We discuss where to apply this framework, opportunities for generalization and future work.

Current scope of framework. We have focused on improving decision making with XAI by generating and testing hypotheses and finding causes. This covers a broad range of explanatory goals. However, Abdul had identified other goals for XAI, such as providing transparency for algorithmic accountability and detecting model bias [2]. Our current model does not address these aspects which relate more towards providing *justification* explanations, to explain if a decision is good [97], has good past

performance [37], or has a trustworthy or authoritative provenance [78].

Miller also identified how explanations are used in social contexts [79]. To accommodate this, we could extend our framework to consider social aspect of explanation, such as argumentation, explanation as cooperative conversation, and dialog [79]. Recent XAI techniques of rationalization or verbalizing explanations by mimicking human generated explanations may be relevant to supporting these social goals (e.g., [26, 62, 89]).

While AI is often used for decision support, deep learning models are being developed primarily for pattern recognition from unstructured data. Essentially, these would serve as intelligent sensors and perform functions, like labeling and counting objects in an image, or recognizing unambiguous human activities (e.g., physical activity, fall detection). Much of our framework on reasoning and decision making may not be relevant for such applications, where users may only seek model transparency and scrutability to debug the model "sensor".

Generalization and extensibility. Our framework provides a starting point to learn how to choose XAI features based on human reasoning. It can be extended by adding specific components to different modules (Section 4.5). For example, we can consider theories of social comparison [31] for health behavior change and connect that with XAI to show prototypes of other users of a fitness app. The user may be persuaded to increase their physical activity if she sees a prototype explanation of users who achieved a health goal. Seeing a counterfactual explanation highlighting the contrast between those prototypes and the user self can allow the user to see what specific lifestyle changes she may have to make.

8 CONCLUSION

We have described a theory-driven conceptual framework for designing explainable facilities by drawing from philosophy, cognitive psychology and artificial intelligence to develop user-centric explainable AI (XAI). Using this framework, we can identify pathways for how specific 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 requirements of human reasoning, we aim to help developers build more user-centric explainable AI-based systems.

ACKNOWLEDGMENTS

We thank Drs. Loh Tze Ping, Ngiam Kee Yuan, Ling Zheng Jye, Adrian Kee, Manjari Lahiri, Loretta Wong, Maria Houdmont, Profs. Sameer Singh, Roland Yap, and Ms Ashley Si for assistance in discussion, prototype design and participant recruitment. This work was carried out in part at the NUS N-CRiPT and BIGHEART research centers and funded by the National Research Foundation, Singapore, Ministry of Education, Singapore, and Google for the APRU-Google Project "AI for Everyone".

REFERENCES

[1] Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications*, 7(1), 39-59.

[2] 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.*

[3] Adams, I. D., Chan, M., Clifford, P. C., Cooke, W. M., Dallos, V., de Dombal, F. T., Edwards, M. H., Hancock, D. M., Hewett, D. J., & McIntyre, N. (1986). *Computer aided diagnosis of acute abdominal pain: A multi-center study. British Medical Journal, 293(6550)*, 800–804.

[4] Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175-185.

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

[6] Arocha, J. F., Wang, D., & Patel, V. L. (2005). Identifying reasoning strategies in medical decision making: a methodological guide. *Journal of biomedical informatics*, *38*(2), 154-171.

[7] 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.

[8] 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.

[9] Barber, D. (2012). Bayesian reasoning and machine learning. Cambridge University Press.

[10] Barnett, G. O., Cimino, J. J., Hupp, J. A., & Hoffer, E. P. (1987). DXplain: an evolving diagnostic decision-support system. *Jama*, *258*(1), 67-74.

[11] Baron, J. (2000). *Thinking and deciding*. Cambridge University Press.

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

[13] Breese, J. S., Heckerman, D., & Kadie, C. (1998, July). Empirical analysis of predictive algorithms for collaborative filtering. *In Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence* (pp. 43-52). Morgan Kaufmann Publishers Inc..

[14] Buchanan, B. G., & Shortliffe, E. H. (1984). Explanation as a topic of AI research. Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project, 331.

[15] 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.

[16] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). 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* (pp. 1721-1730). ACM.

[17] Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

[18] 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.

[19] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

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

[21] Croskerry, P. (2009). Clinical cognition and diagnostic error: applications of a dual process model of reasoning. *Advances in health sciences education*, *14*(1), 27-35.

[22] Croskerry, P. (2017). A Model for Clinical Decision-Making in Medicine. *Medical Science Educator*, *27*(1), 9-13.

[23] Crowley, R. S., Legowski, E., Medvedeva, O., Reitmeyer, K., Tseytlin, E., Castine, M., ... & Mello-Thoms, C. (2013). Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education*, 18(3), 343-363.

[24] Datta, A., Sen, S., & Zick, Y. (2016, May). Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on* (pp. 598-617). IEEE.

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

[26] Ehsan, U., Harrison, B., Chan, L., & Riedl, M. (2018). Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations, AAAI/ACM Conf. on Artificial Intelligence, Ethics, and Society (AIES), 2018.

[27] Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018, March). Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces* (pp. 211-223). ACM.

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

[29] Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018, April). Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 432). ACM.

[30] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115.

[31] Festinger, L. (1954). A theory of social comparison processes. *Human relations*, 7(2), 117-140.

[32] 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.

[33] Guba, E. G., & Lincoln, Y. S. (1982). Epistemological and methodological bases of naturalistic inquiry. ECTJ, 30(4), 233-252.

[34] Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., & Giannotti, F. (2018). Local Rule-Based Explanations of Black Box Decision Systems. *arXiv preprint arXiv:1805.10820*.

[35] Hamblin, C. L. (1970). fallacies. London: Methuen.

[36] Harutyunyan, H., Khachatrian, H., Kale, D. C., & Galstyan, A. (2017). Multitask learning and benchmarking with clinical time series data. *arXiv preprint arXiv:1703.07771*.

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

[38] 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.

[39] Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. Psychological review, 93(1), 75.

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

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

[42] 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.

[43] Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. arXiv preprint arXiv:1712.09923.

[44] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6), 417.

[45] Johnson, A. E., Pollard, T. J., Shen, L., Li-wei, H. L., Feng, M., Ghassemi, M., ... & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific data*, *3*, 160035.

[46] Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow* (Vol. 1). New York: Farrar, Straus and Giroux.

[47] Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H. P. (2018). A cti v is: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, *24*(1), 88-97.

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

[49] 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)*.

[50] 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).

[51] Klein, G. (2007). Performing a project premortem. *Harvard Business Review*, *85*(9), 18-19.

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

[53] Koesten, L. M., Kacprzak, E., Tennison, J. F., & Simperl, E. (2017, May). The Trials and Tribulations of Working with Structured Data:-a Study on Information Seeking Behaviour. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (pp. 1277-1289). ACM.

[54] Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*.

[55] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, (8), 30-37.

[56] 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.

[57] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *In Advances in neural information processing systems* (pp. 1097-1105).

[58] Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W. K. (2013, September). Too much, too little, or just right? Ways explanations impact end users' mental models. In *Visual Languages and Human-Centric Compxuting (VL/HCC), 2013 IEEE Symposium on* (pp. 3-10). IEEE.

[59] 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.

[60] Lakkaraju, H., Bach, S. H., & Leskovec, J. (2016, August). Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1675-1684). ACM.

[61] 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.

[62] Lei, T., Barzilay, R., & Jaakkola, T. (2016). Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

[63] 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), 1350-1371.

[64] Lighthall, G. K., & Vazquez-Guillamet, C. (2015). Understanding Decision-Making in Critical Care. *Clinical medicine & research*, cmr-2015.

[65] Lim, B. Y., Dey, A. K., & Avrahami, D. (2009, 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.

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

[67] 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.

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

[69] Lim, B. Y., & Dey, A. K. (2011, 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.

[70] 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.

[71] Lipton, P. (1990). Contrastive explanation. Royal Institute of Philosophy Supplements, 27, 247-266.

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

[73] Lombrozo, T. (2006). The structure and function of explanations. Trends in cognitive sciences, 10(10), 464-470.

[74] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NIPS 2017)*. (pp. 4765-4774).

[75] Lundberg, S. M., Erion, G. G., & Lee, S. I. (2018). Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888*.

[76] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[77] Markie, P. (2004). Rationalism vs. empiricism. *The Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/rationalism-empiricism. Retrieved 10 September 2018.

[78] McGuinness, D. L., Ding, L., Da Silva, P. P., & Chang, C. (2007, July). PML 2: A Modular Explanation Interlingua. In *ExaCt* (pp. 49-55).

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

[80] 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. *arXiv preprint arXiv:1802.00682*.

[81] 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.

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

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

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

[85] Quinlan, J. R. (2014). C4. 5: programs for machine learning. Elsevier.

[86] 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.

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

[88] Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). 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).

[89] Rosenthal, S., Selvaraj, S. P., & Veloso, M. M. (2016, July). Verbalization: Narration of Autonomous Robot Experience. In *IJCAI* (pp. 862-868).

[90] 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.

[91] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). Learning internal representations by error propagation (No. ICS-8506). *California Univ San Diego La Jolla Inst for Cognitive Science*.

[92] Shortliffe, E. H., & Axline, S. G. (1975). Computer-Based Consultations in Clinical Therapeutics: Explanation and Rule Acquisition Capabilities of the MYCIN.

[93] 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.

[94] Souillard-Mandar, W., Davis, R., Rudin, C., Au, R., Libon, D. J., Swenson, R., ... & Penney, D. L. (2016). Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. Machine learning, 102(3), 393-441.

[95] Sternberg, R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Lawrence Erlbaum.

[96] Sunstein, C. R. (1993). On analogical reasoning. Harvard Law Review, 106(3), 741-791.

[97] Swartout, W. R. (1983). What Kind of Expert Should a System Be? XPLAIN: A System for Creating and Explaining Expert Consulting Programs. Artificial Intelligence, (21), 285-325.

[98] Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, *22*(4-5), 399-439.

[99] Toulmin, S. E. (1958). The Uses of Argument, by Stephen Edelston Toulmin,... University Press.

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

[101] Veale, M., Van Kleek, M., & Binns, R. (2018, April). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (p. 440). ACM.

[102] Vermeulen, J., Luyten, K., van den Hoven, E., & Coninx, K. (2013, April). Crossing the bridge over Norman's Gulf of Execution: revealing feedforward's true identity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1931-1940). ACM.

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

[104] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., & Manzagol, P. A. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of machine learning research, 11(Dec), 3371-3408.

[105] Von Neumann, J., & Morgenstern, O. (2007). Theory of games and economic behavior (commemorative edition). Princeton university press.

[106] Weirich, P. (2008). Causal decision theory. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/decision-causal. Retrieved 10 September 2018.

[107] Whewell, W. (1989). Theory of scientific method. Hackett Publishing.

[108] Zhang, Q., & Li, H. (2007). MOEA/D: A multiobjective evolutionary algorithm based on decomposition. *IEEE Transactions on evolutionary computation*, *11*(6), 712-731.

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