Question-Driven Design Process for Explainable AI User Experiences

Q. VERA LIAO, IBM Research AI, USA

MILENA PRIBIĆ, IBM Design, USA

JAESIK HAN, IBM Design, USA

SARAH MILLER, IBM Research, USA

DABY SOW, IBM Research, USA

A pervasive design issue of AI systems is their explainability–how to provide appropriate information to help users understand the AI. The technical field of explainable AI (XAI) has produced a rich toolbox of techniques. Designers are now tasked with the challenges of how to *select* the most suitable XAI techniques and *translate* them into UX solutions. Informed by our previous work studying design challenges around XAI UX, this work proposes a design process to tackle these challenges. We review our and related prior work to identify requirements that the process should fulfill, and accordingly, propose a *Question-Driven Design Process* that grounds the user needs, choices of XAI techniques, design, and evaluation of XAI UX all in the *user questions*. We provide a mapping guide between prototypical user questions and exemplars of XAI techniques, serving as boundary objects to support collaboration between designers and AI engineers. We demonstrate it with a use case of designing XAI for healthcare adverse events prediction, and discuss lessons learned for tackling design challenges of AI systems.

1 INTRODUCTION

Artificial intelligence (AI) technologies, especially those using Machine Learning (ML) algorithms, have become ubiquitous even in high-stakes domains such as healthcare, finance, business and healthcare. How to make AI explainable, that is, providing appropriate information to help users *understand* its functions and decisions, has been considered one of the critical and pervasive design issues for AI systems [10, 39]. This understanding is a necessary foundation to serve many user needs such as to improve, contest, develop appropriate trust and better interact with AI, and eventually accomplish the user goal, whether it is to make better decisions or to delegate tasks to AI with confidence. However, at the present time, explainability features are still not a common presence in consumer AI products. Recent studies, including our own work [39], looking into industry practices around explainable AI paint a dim picture where while "explainability was strongly advised and marketed by higher-ups" [15], product teams are grappling with tremendous challenges to implement the right solutions, and more broadly, a lack of frameworks to decide when, where and how to provide AI explanations.

The time is ripe to start tackling these challenges and establish a shared understanding and best practices on how to design explainable AI systems. In the past few years, the academic field of explainable AI (XAI) has seen a fruitful surge, thanks to the growing public interest in fair, transparent and responsible AI. Hundreds of, if not more, algorithms that produce various types of explanations have been proposed in the literature [8, 11, 30]. Meanwhile, many open source libraries and toolkits of XAI [1, 3–5] emerged recently, making XAI algorithms originated in the academic community into practitioners' toolbox.

The availability of a toolbox does not guarantee its intended utilization and benefits. First, with a toolbox there is the challenge of selection–how to effectively and efficiently identify a suitable XAI technique for a given application or interaction when there are an abundance of, and rapidly growing, choices. Second, there is the challenge of translation– how to make technical innovation made by academic research effective and usable in real-world sociotechnical systems. This problem is especially prominent for XAI, as the technical XAI field has been criticized for developing algorithms in a vacuum [44]. Many XAI algorithms were developed with the intention of allowing scrutiny of ML models during model developing or debugging process. So the intended users are often ML engineers or researchers, whose abilities to understand AI and reasons to seek explanations may differ remarkably from end users of consumer AI products. To utilize these XAI algorithms in diverse AI products will inevitably require adaption of the algorithms, translation of the algorithmic output, and effectively integrating them in the overall user workflow and experiences.

Designers and UX researchers of AI products are often burdened with these challenges. Recent studies [26, 39] looking into design processes of XAI systems break down the design challenges into two parts. One is to answer *what to explain*, i.e., to identify the right explanation content based on product, user, context and interaction specific needs. Second is to address *how to explain*, to create the right presentation format for the explanation, which often requires iterative design and evaluation. These two stages can be seen as corresponding to the challenges of selection and translation to use the toolbox of XAI techniques.

To tackle the design challenges of XAI also requires tackling general challenges that designers face working with AI systems, which have received growing attention in the human-computer interaction (HCI) and design communities. One challenge is the need to form a "designerly" understanding of what AI can or cannot do to be able to envision possible interactions [25, 29, 59]. This is specially difficult in the context of designing XAI because designers may need to understand not one, but a toolbox of XAI algorithms. This is also especially necessary for XAI because current XAI algorithms cover only a subset of the landscape of explanations in human experiences [44, 45]. What kind of AI explanations can be provided to users is often constrained by the technical affordance. As our previous study looking into design practices around XAI [39] found, the critical challenge for XAI UX is "finding the right pairing to put the ideas of what's right for the user together with what's doable given the tools or the algorithms".

The second challenge is for designers to be able to collaborate effectively with AI engineers (both broadly defined[1]) [29, 59]. This is a unique problem to AI product teams, as the design goals should be formulated with a shared vision with AI engineers based on the technical viability, and ideally impact the model development and evaluation. However, in reality AI experts are a scarce resources and designers are often found in a challenging situation to collaborate with them due to a lack of common knowledge and shared workflow. When it comes to designing XAI UX, our previous study found that communication barriers preventing buy-in from data scientists, who are burdened with implementing XAI techniques, are major roadblockers for designers to advocate for AI explainability [39].

---

[1]For simplicity, for the rest of the paper, we refer all roles that perform design tasks as "designer", which may also include UX researchers, user researchers, etc. We refer all roles that perform engineering tasks on the model as "AI engineer", which may include data scientists, AI researchers, etc.

Recent work suggests that AI may be "a new and difficult design material" [25, 58] that requires developing AI specific user-centered design processes. The process should help designers form a high-level understanding of AI's technical capabilities as they explore the problem and solution space of their design, as well as closer and iterative collaborations with AI engineers instead of treating design as the final deliverable. We further argue that it is necessary to develop AI design-problem specific user centered design processes, which, besides explainability, include fairness, robustness, errors and uncertainty, among others. These are known design issues that are universally present in AI systems, and critical to AI UX that a product team may choose to dedicate their design effort on. Similar to XAI, each of these topics has a vast collection of algorithmic solutions contributed by an an active academic community, and made increasingly accessible by open-source toolkits. When treating algorithmic solutions as "a toolbox of design materials", the challenges of selection and translation are likely shared problems.

In this work, we contribute a novel design process to enable designers and product teams to work with a toolbox of AI algorithms, focusing on the design issue of AI explainability. We start by reviewing related work on XAI and AI design, and identify four requirements that a design process for XAI UX should satisfy. Based on these requirements, and building upon prior work on designing explainable technologies, we propose a *Question-Drive Design Process for XAI*, which grounds identification of user needs and requirements, choices of XAI techniques, design and evaluation of XAI UX solutions, all in the *user questions* to understand the AI. By mapping user questions to exemplars of XAI algorithms, for which we provide a mapping guide, this process also enables the user questions to serve as boundary objects to support consensus building between designers and AI engineers.

Below we first review related work, leading to the requirements that guided our proposal. Then we describe in detail the *Question-Driven Design Process for XAI UX*, and demonstrate it with a case study of designing XAI for healthcare adverse events prediction following this process. Lastly, we discuss considerations and feedback for practitioners to adopt this design process, and broader implications for designing AI UX.

2 RELATED WORK

2.1 Explainable AI

While the origin of explainable AI could be traced back to earlier work in expert systems [51], the re-surge of this wave of XAI work could be attributed to an increasing popularity of opaque, difficult-to-understand machine learning (ML) models such as those using deep neural networks. The academic research of XAI was accelerated by legal requirements such as EU's General Data Protection Regulation (GDPR) and various government funding programs [31], as well as broad industry awareness for explainable and responsible AI. In short, the field has seen booming efforts in the past five years or so, developing a vast collection of new algorithms.

While there is no agreed-upon definition or criteria of AI explainability, and there are different related terms used in the literature such as interpretability and transparency, a common goal shared by XAI work is to make AI's decisions or functions *understandable* by people [8, 11, 12, 30, 41, 44, 45]. This understanding can then support various goals for people to seek explanations, such as to assess AI capability, trustworthiness and fairness, to diagnose and improve the model, to better control or work with AI, and to discover new knowledge [8, 39, 44]. This wave of XAI work has focused dominantly on ML models, although explainbility of other types of AI systems such as planning [20], multi-agent systems [48], etc. are receiving increasing attention. Following this trend, this paper will primarily focus on designing explianable ML systems.

A complete overview of the landscape of XAI algorithms is beyond the scope of this paper. Multiple recent papers surveyed this field and converged on a few important dimensions to classify and characterize XAI algorithms [11, 12, 30]. One is the differentiation between directly explainable models and opaque models, the latter of which often require using separate algorithms to generate *post-hoc* explanations. In terms of the scope of explanation, XAI work differentiates between global explanations, which provide a general understanding of how the model works, and local explanations, which focus on explaining how the model makes a prediction for a particular instance. Guidotti et al. [30] further differentiates a third category: model inspection, which focuses on explaining a specific property of a model such as how changes of a feature impact the model prediction. More fine-grained classification methods consider the formats of the explanation, which could range from less than 10 [30] to more than 40 types [11], each of which could be implemented by multiple algorithms that differ in computational properties.

In short, the technical field of XAI have produced hundreds of algorithms that produce various forms of explanation. Recently, open-source toolkits such as IBM AIX 360 [3], Microsoft InterpretML [4], H2o.ai Machine Learning Interpretability [1] are making popular XAI algorithms (e.g. [22, 42, 46, 47]) increasingly accessible to practitioners. It has become necessary to develop methods that could help practitioners identify XAI algorithms suitable for their purposes. Such methods should be actionable for practitioners to follow, and scalable as the field is still rapidly advancing.

2.2 HCI and user-centered perspectives of XAI

The HCI community have a long-standing interest in algorithmic transparency and making computing systems explainable. Most notably, a 2018 paper by Abdul et al. [7] conducted a literature analysis of HCI research on explainable systems, ranging from expert systems, recommenders, context-aware technologies to more recent ML systems. The authors identified opportunities for HCI research agenda to focus on improving the rigor, by testing effectiveness of explanation interfaces in real-world systems, and improving the usability, by drawing from rich body of cognitive science, design and HCI knowledge on how people consume explanations. Similar views have emerged in the AI community pushing towards a human-centered instead of algorithmic centered XAI field, calling out gaps between XAI algorithmic output and properties of explanations sought by people [44], and the necessity to produce and evaluate AI explanation by involving the intended user group in the intended usage context [24].

HCI researchers have begun designing explainable AI systems by utilizing various XAI algorithms and making them usable. Some studied explainable interfaces to help model developers diagnose and improve ML models [34, 35]. Others studied explainability features to help decision-makers such as physicians [54, 56] and recruiters [21] to better understand AI recommendations and make informed decisions. It is common for these systems to include multiple XAI features, as users often demand different kinds of explanations at different usage points. Novel usage of AI explanations have also been explored. For example, Lai et al. [37] leveraged XAI algorithms to generate rationales of toxicity detection as tutorials to train people to perform content moderation. Ghai et al. [28] proposed explainable active learning, which utilizes AI explanations as interfaces for annotators to provide feedback to train models. Another active research area focuses on rigorously evaluating the effectiveness of XAI techniques [9, 18, 19, 23, 57]. For example, Dodge et al. [23] compared the effectiveness of multiple popular forms of XAI techniques in supporting people's fairness judgment of ML models, and found that there is no one-fits-all solution.

These HCI studies have generated valuable insights for tackling the challenges of selection and translation of the toolbox of XAI algorithms. They often suggest domain or user group specific guidelines on what kind of XAI technique works effectively, and what interface features should be provided. In a parallel effort, recent work proposed several taxonomies of prototypical roles of XAI consumers [11, 33, 52], as a starting point to provide role-based guidelines to

choose between XAI techniques. Typical user roles include model builders, decision-makers, people being impacted by AI, business owners, regulatory bodies, etc.

While these efforts can be seen as producing top-down guidelines, recent studies begun to explore bottom-up, user-centered approaches to designing XAI UX [26, 39, 55]. This paper is built upon our previous work [39] studying design challenges around XAI by interviewing 20 designers working across 16 AI products. We identified a key challenges for creating XAI UX is the variability of users' explainability needs, i.e. types of explanation sought. Based on prior HCI work on "intelligibility types" [40] and social science literature on explanatory relevance [32] that defines different types of explanation, we argued to represent user needs for explainability by the type of question the user asks to understand the AI, such as *Why*, *What if*, *How*, etc. Our study found that many factors could vary the type of question a user asks, including their motivation to seek explanation, points in a user journey, model type, decision contexts, among others. This observation highlights that guideline-based approaches may not be granular enough to allow system, user, and interaction specific selection and design of XAI techniques, and thus a user-centered approach to ground the design in a nuanced understanding of users and the usage contexts is often necessary. Furthermore, in that study we "crowdsourced" common user questions from the designers and created an XAI Question Bank, shown in Figure 1 (with minor updates from [39]). The questions are organized by a taxonomy of 9 categories, representing prototypical questions people have for different types of AI explanation, focusing on ML systems. Following this effort, in this work we will propose a user-centered design process for XAI that starts by identifying the types of explanation needed by eliciting user questions. The proposed design process will also use the XAI Question Bank (Figure 1) as an auxiliary tool.

2.3 What's unique about designing UX of AI

Our proposed method is also motivated by recent work looking into the unique challenges of designing user experiences for AI systems. A common challenge is the need for designers to understand ML capabilities [25, 58, 59]. Interestingly, prior work reveals that the required is often not technical knowledge of ML but a "designerly" understanding of what a ML technique can or cannot do, and what user values it can deliver. According to Yang et al. [59], designers often rely on abstraction of ML capabilities and exemplers of existing AI systems to establish such a designerly understanding. A related area of research that could support a designerly understanding of and formulating design goals for AI aims to develop design guidelines specific to AI systems. Notable examples include Amerishi et al.'s heuristic guidelines for human-AI interaction [10], and Madaio et al.'s checklist for AI fairness [43].

Another area of challenges arise from the unique design process of AI systems. This process often takes longer than non-AI products and should ideally involve designers and AI engineers in a tight collaboration from concept development, co-evolving of design and technical development, and continued refinement after deployment [59]. However, there is currently no formal requirement or best practices for designers and AI engineers to collaborate. Designers' interactions with AI engineers are often ad-hoc, sporadic and no shortage of communication challenges. By reflecting on the lifecyle of how data scientists build ML models and designers' work process, Giradin and Lathia [29] identified a series of touch points for designers to work with data scientists to produce meaningful UX for AI powered systems, including co-creating a tangible vision of experience with priorities, goals and scope, assessing the assumption of algorithms with insights from user research, articulating solutions and understanding the limitations from both sides, and specifying successful metrics of the user experience that may impact choices of data and algorithms.

This challenge for designers to work with AI engineers was also foregrounded in our previous study on design challenges around XAI [39]. Because providing explanability features often requires AI engineers to implement either additional algorithms or functions to generate information about the model, communication barriers and failure to

XAI Question Bank

Data
- **What kind of data was the system trained on?**
- What is the source of the training data?
- How were the labels/ground-truth produced?
- What is the sample size of the training data?
- What dataset(s) is the system NOT using?
- What are the limitations/biases of the data?
- What is the size, proportion, or distribution of the training data with given feature(s)/feature-value(s)?

Output
- **What kind of output does the system give?**
- What does the system output mean?
- What is the scope of the system's capability? Can it do…?
- How is the output used for other system component(s) ?
- How can I best utilize the output of the system?
- How should the output fit in my workflow?

Performance
- **How accurate/precise/reliable are the predictions?**
- How often does the system make mistakes?
- In what situations is the system likely to be correct/ incorrect?
- What are the limitations of the system?
- What kind of mistakes is the system likely to make?
- Is the system's performance good enough for…?

How (global)
- **How does the system make predictions?**
- What features does the system consider?
  - Is [feature X] used or not used for the predictions?
- What is the system's overall logic?
  - How does it weigh different features?
  - What kind of rules does it follow?
  - How does [feature X] impact its predictions?
  - What are the top rules/features that determine its predictions?
- What kind of algorithm is used?
  - How were the parameters set?

Why
- **Why/how is this instance given this prediction?**
- What feature(s) of this instance determine the system's prediction of it?
- Why are [instance A and B] given the same prediction?

Why not
- **Why/how is this instance NOT predicted to be [a different outcome Q]?**
- Why is this instance predicted [P instead of a different outcome Q]?
- Why are [instance A and B] given different predictions?

How to be that
- **How should this instance change to get a different prediction Q?**
- What is the minimum change required for this instance to get a different prediction Q?
- How should a given feature change for this instance to get a different prediction Q?
- What kind of instance is predicted of [a different outcome Q]?

How to still be this
- **What is the scope of change permitted for this instance to still get the same prediction?**
- What is the range of value permitted for a given feature for this prediction to stay the same?
- What is the necessary feature(s)/feature-value(s) present or absent to guarantee this prediction?
- What kind of instance gets the same prediction?

What If
- **What would the system predict if this instance changes to…?**
- What would the system predict if a given feature changes to…?
- What would the system predict for [a different instance]?

Others
- How/why will the system change/adapt/improve/drift over time? (change)
- Can I, and if so, how do I, improve the system? (change)
- Why is the system using or not using a given feature/rule/data? (follow-up)
- What does [a machine learning terminology] mean? (terminological)
- What are the results of other people using the system? (social)

Fig. 1. XAI Question Bank for explaining supervised machine learning systems we developed in [39], with minor updates

reach a shared design vision often impede buy-in from AI engineers, especially when explainability is at odds with other product goals such as time pressure to deliver minimum viable products (MVPs).

Other design challenges of AI UX discussed in the literature include a shift towards data-driven design culture [17, 36], challenges to prototype and engage in quick design iterations [13, 50, 53, 59], and the needs to engage stakeholders to align the values of AI system [38, 60]. Our work will primarily tackle supporting a "designerly" understanding on the technical space of XAI, and designer-AI engineer collaboration, as discussed in the requirements below.

3 QUESTION-DRIVEN DESIGN FOR EXPLAINABLE AI USER EXPERIENCES

We first discuss the requirements that guided our proposal, and the method we follow to develop . Then we provide an overview of the design process, and describe a use case in which we introduced this design process to a team working on an AI system that provides healthcare adverse event predictions. We elaborate on each stage of the design process, together with the design activities we carried out working with the use case and their outcomes.

3.1 Requirements and Development Method of the Design Process

Based on related work reviewed above, we set out to propose a design process for creating XAI UX with the following requirements:

- **R1:** *Following a user-centered approach to identify user, context and interaction specific needs for AI explainability.* Following our prior work [39], we ground user needs, specifically what type of explanation is sought, in the type of *question* that a user asks to understand the AI.

- **R2:** *Enabling a designerly understanding on the affordance of different XAI techniques available.* We do so by mapping prototypical user questions to candidate XAI technical solutions, focusing on ones that are available in popular open-source toolkits, and providing exemplars of explanation they could generate (in verbal descriptions or visual examples). In other words, this mapping explicitly links existing XAI techniques with values they could provide to users, as grounded in the type of user question they could answer.

- **R3:** *Supporting collaborative problem solving between designers and AI engineers to identify XAI UX solutions that are both technically viable and satisfying user needs.* Specifically, the user questions can be used as *boundary objects* to facilitate the communication and consensus building, in which designers can ground their understanding on user requirements for asking a question, and AI engineers can ground their understanding on the technical details, supported by the mapping guide mentioned above.

- **R4:** *Supporting an end-to-end, iterative design process by articulating the design goals or success metrics of XAI to guide continued, in-parallel refinement of design and technical solutions.*

With these requirements in mind, we developed Question-Driven Design Process for XAI first through a series of ideation workshops with a group of 11 volunteers, consisting of designers, data scientists and XAI researchers from multiple AI product lines of a large international technology company. We then presented the initial idea of the design process using a slideshow to an additional 10 designers, HCI researchers and data scientists for feedback, and iterated on the content.

So far the design process has been carried out in two use cases for validation and refinement. One is an end-to-end design process for a team to develop an explainable AI system that provides healthcare adverse event prediction to help healthcare providers identify high-risk patients and improve their care management. We will use this use case to illustrate the design process in the following sections. Another use case is an ongoing project to develop explainable AI solutions for crime analytics. So far the team has adopted the first half of the design process to understand user needs and requirements for XAI. It provides a more complex use case where users are not just receivers of AI predictions but also need to configure their own models. This use case helped us refine the instruction of the design process to make it flexible for different teams' needs, which we will reflect on in later sections.

3.2 Design Process Overview

Figure 2 provides an overview of the design process, consisting of four steps: question elicitation, question analysis, mapping questions to modeling solutions, and iterative design and evaluation. In the forth row we suggest the stakeholders that should be involved in each step. As we sought feedback for the design process from designers and AI engineers, we became aware that AI product teams differ significantly in their ability to access users, availability of AI experts, designers' as well as the team's general AI expertise, and the nature of the AI technology. Therefore, in the last row we suggest possible workarounds and additional resources if a product team face difficulty executing a step, which we will discuss as *supplementary guidance* in later sections.

We also provide in the supplementary material of this paper a slideshow for the design process, with which we are introducing this design process to product teams by hosting workshops and as an introductory handout. In addition to the detailed description of the design process and the use case to be described, the slideshow also includes an overview

	Step 1 **Question Elicitation**	Step 2 **Question Analysis**	Step 3 **Mapping Questions to Modeling Solutions**	Step 4 **Iterative Design and Evaluation**
Goals	• Elicit user needs for explainability as questions • Gather user intentions and expectations for the questions	• Categorize elicited questions and identify priorities • Identify key user requirements for the XAI UX	• Map prioritized question categories to candidate XAI technical solutions as a set of functional elements that the design should cover	• Create a design including candidate solutions and evaluate it with the user requirements • Iteratively improve the design to close the gaps
Tasks	1. In user research, first define the problem the AI solves, the AI tasks and/or user journey 2. Elicit questions that the AI needs to answer in the tasks or user journey to be useful, effective and trustworthy 3. Also articulate the intentions behind asking the questions and expectations for the answers	1. Cluster similar questions across users into categories 2. Prioritize categories that the XAI UX should focus on. This can be done by ranking the quantity of questions or collaboratively with the team 3. Summarize and cluster user intentions and expectations to identify key user requirements for the XAI UX	Work with AI engineers to collaboratively identify candidate technical solutions to generate explanations for each prioritized question category. A mapping guide to available techniques in XAI toolkits is provided. Sometimes it requires implementing solutions specific to your model	1. Create an initial design with the set of candidate XAI solutions 2. Evaluate the design holistically with the user requirements identified in step 2, ideally with user feedback 3. Iteratively improve the design and evaluate with the user requirements to close the gaps 4. If applicable, define success metrics based on the user requirements to guide the improvement of models and XAI techniques
Stakeholders involved	Designers, users	Designers, optionally the product team	Designers, AI engineers	Designers, AI engineers, preferably users
Supplementary guidance	• When access to users is limited, designer can use XAI Question Bank as a checklist to guide the elicitation. Designer can also adapt it to create a customized checklist • The team can also use the checklist to heuristically identify applicable user questions • A complex AI system may need to be broken down into steps or scenarios to elicit questions separately • For highly novel systems, scenarios or low-fi prototypes can be used to elicit questions	• For supervised ML, XAI Question Bank can be used to guide the categorization into 9 categories. However, categorization can be flexible based on the questions elicited • The analysis offers an opportunity for the team to develop a shared vision, so it could be beneficial to involve the team to vote on the user questions as team priorities, for example by using a prioritization matrix	• Depending on the XAI expertise level of the team, additional training or educational materials may be needed before scheduling this step (suggestions in the supplementary material) • Depending on the product stage, this step could impact not only the choice of XAI techniques but also the model (e.g., a simpler, more explainable model)	• In addition to the mapping guide, AI engineers can provide additional exemplars of output of XAI techniques to support design ideation • When access to users is limited, user requirements can be used for a heuristic evaluation within the team • Design iterations often impact, and should be developed in parallel with modeling solutions • New questions, user requirements and additional XAI techniques can emerge and should be incorporated in future iterations

Fig. 2. Overview of Question-Driven Design Process for XAI UX

of what XAI is and more detailed exemplars of XAI techniques with visual illustrations. In the Discussions section, we will reflect on the lessons learned on how to introduce a new design process to product teams.

The first two steps, question elicitation and analysis, are to enable designers and the team to understand the user needs and requirements for the XAI UX. The last two steps, mapping modeling solution and iterative design and evaluation, are for designers and AI engineers to engage in collaborative problem solving for technology and UX co-development based on user needs and requirements.

Based on **R2** identified above, the collaborative problem-solving could be accelerated by supporting a designerly understanding on the affordance of different XAI techniques. We do so by mapping user question to candidate solutions of XAI techniques, with example descriptions of their output. In practice, we observed that currently AI engineers in product teams are often not familiar with various XAI algorithms. So we also link these candidate solutions to readily usable toolkits or code libraries to support AI engineers' technical understanding of XAI solutions and accelerate the implementation, which could help fulfilling **R3**.

In Figure 3, we provide a suggested mapping guide between user questions and descriptions of candidate technical solutions (in the slideshow we also provide visual exemplars), as well as links to code libraries. These questions are based on the categories in the XAI Question Bank [39] (Figure 1). Most links are for post-hoc XAI algorithms, i.e., they can be used to generate explanations for complex models that are not directly explainable. When using simpler models

Question	Explanations	Example XAI techniques
Global how	• Describe what algorithm is used and what features are considered, if a user is only interested in a high-level view • Describe the general model logic as feature impact*, rules+ or decision-trees• (sometimes need to explain with a surrogate simple model)	ProfWeight*, Feature Importance*, PDP*, BRCG+ , GLRM+ , Rule List+ , DT Surrogate•
Why	• Describe what key features of the particular instance determine the model's prediction of it* • Describe rules+ that the instance fits to guarantee the prediction • Show similar examples• with the same predicted outcome to justify the model's prediction	LIME*, SHAP*, LOCO*, Anchors+, ProtoDash•
Why not	• Describe what changes are required for the instance to get the alternative prediction and/or what features of the instance guarantee the current prediction* • Show prototypical examples+ that had the alternative outcome	CEM* , Prototype counterfactual+ , ProtoDash+ (on alternative class)
How to be that	• Highlight features that if changed (increased, decreased, absent, or present) could alter the prediction* • Show examples with small differences but had a different outcome than the prediction+	CEM*, Counterfactuals*, DiCE+
What if	• Show how the prediction changes corresponding to the inquired change	PDP, ALE, What-if Tool
How to still be this	• Describe feature ranges* or rules+ that could guarantee the same prediction • Show examples that are different from the particular instance but still had the same outcome	CEM*, Anchors+
Performance	• Provide performance metrics of the model • Show confidence information for each prediction • Describe potential strengths and limitations of the model	Precision, Recall, Accuracy, F1, AUC Confidence FactSheets, Model Cards
Data	• Document comprehensive information about the training data, including the source, provenance, type, size, coverage of population, potential biases, etc.	FactSheets, DataSheets
Output	• Describe the scope of output or system functions • Suggest how the output should be used for downstream tasks or user workflow	FactSheets, Model Cards

Fig. 3. Mapping guide between question categories and candidate XAI techniques, with exemplary descriptions of explanations provided by these techniques and links to code libraries

such as regression or decision tree, it is possible to generate such explanations directly based on model internals using standard data science packages, as we elaborated in the slideshow. The last three categories are considered "model facts" to describe the model's performance, training data or output. The corresponding links (FactsSheets [6], Model Cards [2] and Datasheets [27]) provide exemplars of industry standards for model documentation. This mapping was built by reviewing various survey papers of XAI techniques [8, 11, 30], and iterated on with feedback from 5 data scientists and XAI researchers. In particular, we focused on suggesting XAI techniques that are available in current XAI toolkits.

We emphasize that this mapping guide is suggestion only and focuses on explaining standard supervised ML technologies, in which users receive predictions or recommendations from AI to assist their decision-making. In practice, it is possible that an AI system could invoke new categories of user question, or a product team may choose to categorize user questions in different ways. It is important for designers and AI engineers to work together to identify model specific solutions. Sometimes the solution may lie outside the scope of this mapping guide but requires generating information about a specific AI model's parameters or internals, for which the team could use standard data science frameworks or implement a new technique.

3.3 Use Case Overview

In 2020, a team in the technology company were working with a large hospital to develop an AI system that predicts patients' risk of healthcare adverse events, for example unplanned hospitalization due to an urgent and unexpected condition. The model prediction is made based on a patient's profile such as medical history and social-demographic information. The intended users are healthcare providers, who could use the AI's predictions to identify potential high-risk patients to better manage their healthcare plans. It was recognized that explainability, to enable healthcare providers to understand the AI and its predictions, would be critical to the user experience. We joined the project team and introduced the question-driven design process, and worked with the team to carry out the process.

The multi-disciplinary team consisted of designers, UX researchers, data scientists, software engineers, XAI researchers, and a project manager. The team conducted user research with 9 participants (doctors and research nurses). 4 of them are sponsored users from the company and 5 are recruited in the hospital. The prediction model is a Deep Learning model based on custom recurrent neural network (RNN) architectures and trained mainly with medicare claim data. The details of the model and data collected in the hospital are considered proprietary and will not be disclosed in this paper (we will exclude quotes from participants recruited in the hospital). Instead, we will focus on presenting the activities and outcomes by following the question-driven design process. We will not delve into the details of the user research as our goal is to illustrate the design process rather than presenting empirical insights from the use case.

3.4 Step 1: Question Elicitation

The first step *Question Elicitation* aims to gather empirical data to understand user requirements for the XAI UX. This step mainly involves designers and targeted users participating in the user research. While it can be part of a broader formative user research using interviews, focus groups, open-ended surveys, etc, we suggest to center the inquiry for explainability around what *questions* users have for understanding the AI. The notion of representing user needs for explainability, or the type of explanation they seek, as questions can be found in HCI literature [39, 40] as well as social sciences literature [32].

The empirical data gathered in this step will be used to fulfill **R1**, to identify what types of explanations are needed for a specific user group, a usage context and or an interaction, and **R4**, to help articulating the design goals of the XAI UX. Therefore, we suggest to engage in two parts of user inquiry. First is to ask participants to come up with questions they have for the AI. A prompt we found useful is "*What questions does the AI need to answer for it to be useful, effective and trustworthy?*" Second is to follow up and ask participants to articulate the intentions behind asking a question, and their expectations for answers to the question. Prompts we found useful include "*Why would you ask this question*" and "*What would a good answer to this question look like?*"

The question inquiry should happen after defining the problem space and the AI tasks, so that participants can effectively engage in such a reflection. The task definition can either be done with the participants as part of the user research, or provided directly to scaffold the discussion.

3.4.1 Activities in the use case. As we conducted formative user research for the AI system for healthcare adverse event prediction, we engaged in question elicitation as part of the user interviews with 9 participants (doctors and research nurses). The whole interviews lasted 45-60 minutes each. The first half focused on a broader understanding of the task of preventing healthcare adverse events, asking questions such as "*How do you usually assess patients' risks*" and "*what do you do to prevent unplanned hospitalization?*" Then we asked participants to envision "*How can an AI system help with adverse events risk prediction and prevention*?" Depending on participants' experience level with AI, we selectively

provided high-level descriptions of AI prediction capabilities and guided the discussion towards envisioning concrete tasks the AI can perform. Participants converged on the vision of an AI system assessing patients' risks or alerting high-risk patients for them to pay attention to, and helping them identify appropriate interventions. At this point, we begun the question elicitation by asking:

- *Imagine using such an AI system, what do you want to understand about it? What questions would you ask?*
- *What questions would you want the AI to answer when you look at its risk prediction for a patient?*
- *Are there any other questions the system should be able to answer for you to trust or feel comfortable using it?*

After participates articulated their questions, we followed up by probing for their intentions to ask these questions, and their expectations for the answers whenever there were non-obvious ones, by asking:

- *Why would you be interested in asking these questions?*
- *What would be a good answer look like? What should the AI tell you?*

All interviews were video recorded and transcribed. Below we focus on the part of data around elicited questions. Sometimes user questions appeared in other parts of the interviews without the elicitation prompts. We also included these questions and discussions around them.

3.4.2 Outcomes in the use case. On average, we elicited 6 questions from each participant, with a maximum of 12 questions and minimum of 3 questions. We observed that participants' experience with AI impacted the quantity and breadth of questions they formulated. For example, P1 is a sponsored user recruited in the company, who has had extensive experience working with AI systems. She came up with 12 questions that were clustered in three areas. The first area is about understanding the AI's risk prediction for a patient:

- *What is the risk the patient has and what can the outcomes be?*
- *Why this risk?*
- *Is there a risk because...[they have not taken the drug as prescribed]?*
- *What are the main risk factors for this person?*
- *What should I be worried about this patient?*

Based on the taxonomy of XAI Question Bank, the first one is an *Output* question, and the last four are *Why* questions. Without us asking, P1 proactively elaborated on the intentions and expectations for asking these questions: "*it would be helpful if the AI could help me really understand the patient...is there a chronic condition, the reasons they were being admitted...should I put them in a risk of mobility or even mortality*"; and: "*...understanding why this risk, ideally in a way that I can either address, or acknowledge that, or make some kind of change*".

Then she moved on to the second area that directly asked about changes she could make to reduce the patient's risk:

- *Is there some kind of adjustment or accommodation I can make for the patient's care?*
- *For patients with the same risk profile, what have been done for the success stories?*
- *What happened to the unsuccessful stories [of patients with the same risk profile] ?*

The first two are *How to be that* (less risk) questions. And the last one is a *How to still be this* question. They were all concerned about seeking appropriate interventions, as P1 followed up to explain the intention: "*Keep[ing] myself honest about what I am going to do [with this patient], whether it is aligned with what has been practiced... guidelines are for broader population, but they are not personalized enough. So it would be better to know what would specifically work for patients like mine.*"

11

After we further probed on what questions would make her feel comfortable using the system, she asked about the algorithm (*Global how*) and *Data*.

– *What is the model or algorithm that is generating the predictions?*
– *Why is this model chosen?*
– *What are the demographics of the population in the training data?*
– *What medical conditions do patients in the training data have?*

She followed up with rich discussions about her intention for asking these question, as mainly to understand the AI's limitations to decide when to confidently rely on it, and the expectation for the explanation, to be both concise and actionable: "*..a little bit of background, or insights on what the AI is, what does it applies to. Just a little bit! ... It might be too much for some people, but it would be nice to have a cheatsheet of the methodology. Sometimes people just say we use deep learning or we use logistical progression or whatever. But why? I just want to know, basically, if I use this version, does it mean I might miss out a little bit of this, or does it take the bigger picture thing... As my own due diligence, understanding why it was chosen, because at the end of the day we are responsible for what this does. Especially when [its prediction] doesn't click with my clinical experience I will have to actually review it and just be like, so this is why so there could be a chance that there's a little off. When it comes to this kind of scenarios, maybe I should lean more of my own experience*"

In summary, a total of 54 questions, and 37 excerpts of discussion on the intention and expectation for asking the questions were collected in Step 1, and will be used for analysis in Step 2.

3.4.3 Supplementary guidance. In the use case, we were able to recruit 9 participants. A common feedback we got is that some projects may have limited access to interviewing users, especially in an agile environment or in the initial concept development stage, so they may not be able to gather a sufficient amount of questions to ensure a thorough understanding of user needs. One workaround, we suggest, is to use XAI Question Bank (Figure 1) as a checklist, asking users to go through them and identify applicable questions. This approach could enable a more comprehensive coverage of the space of XAI with a small number of participants. It is even possible for the the product team to use the checklist to heuristically identify applicable user questions, which could be useful in the early stage of concept development. The XAI Question Bank [39] has 50 questions and intentionally include common variations. Designers can also start with selecting from or tailoring these questions to create a customized checklist.

Another flexible decision point that can be tailored to a design project is the context of question elicitation–how well-defined the AI tasks should be, and how to define them. In our use case, we asked participants to define the AI tasks, because we were essentially introducing an AI system to assist an existing well-defined problem in their job. The interview protocol, by asking them to first reflect on current practices and the problem space, effectively facilitated them to define concrete and mostly consistent AI tasks. However, this is likely more challenging if the intended AI tasks are complex or less familiar. If the AI tasks are ill-defined or vary significantly across participants, the question elicitation may generate less value. We suggest several possible workarounds. First, for a complex AI system, it may need to be broken down into steps or scenarios to elicit questions separately. Providing a user journey map or user scenarios can be useful. Second, for highly novel systems that are hard to come up with questions for, user scenarios or even low-fi prototypes can be used to scaffold the elicitation.

3.5 Step 2: Question Analysis

The second step *Question Analysis* is to analyze the data collected in Step 1, including the questions and content discussing the intentions and expectations. The goal is to establish an understanding on the user needs and requirements to

guide the design and evaluation of XAI UX, both in terms of the types of explanation to provide (**R1**), and key user requirements as design goals (**R4**). This step can be completed by designer(s) who then present the insights to the team, or serve as an opportunity to establish a shared team vision by involving the whole team in analyzing the questions and collaboratively identifying priorities (**R3**).

The first step is to cluster similar questions across participants to arrive at main categories of user questions. For standard supervised ML systems, our XAI Question Bank [39] (Figure 1) suggests main categories and variations of question that could help guide the categorization. Depending on the nature of the AI system and user needs, one may identify new categories of questions that are not presented in XAI Question Bank.

After the categories are established, the task is to identify which categories should be priorities for the XAI UX to focus on. This prioritization task can be approached in multiple ways. One is to simply count the number of questions under each category, thus the occurrences of such needs across different users, and prioritize those asked most frequently. Or a team can review the questions and decide on the priorities collaboratively

Another area of the analysis is to identify main themes from participants' comments on their intentions and expectations for asking the questions, as key user requirements for XAI UX. This can be done by following standard qualitative analysis approach, by first summarizing or tagging individual excerpts and then clustering similar ones to identify emerging themes. These user requirements will then be used to guide the evaluation of the XAI UX in later steps. We highlight that there is often not a one-to-one mapping between questions and intentions. Participants may ask different questions for the same underlying reason, or similar questions but for different reasons. So we recommend to analyze and cluster the user requirements separately. It may also be necessary to refer to other parts of the interview, for example discussions about the user background or workflow, to better interpret the comments about why or what a user needs by asking a question.

3.5.1 Activities in the use case. Two authors of the paper worked with designers to perform the above-mentioned analyses and presented the insights to the team. Since adverse events prediction is a standard supervised ML regression task, we found the questions largely aligned with the taxonomy provided in the XAI Question Bank. However, a challenge we noticed is the need to translate questions to the general forms appeared in the XAI Question Bank. One reason is that the healthcare domain has its own familiar jargon, for example using the terms "sensitivity and specificity" instead of standard ML performance metrics. Another reason is that participants often asked questions based on their workflow, in a more specific way than the form appeared in the XAI Question Bank. For example, for a *Why* question about understanding the risk factors (features) of a patient, participants asked "*What comobility factors does this patient have [to make him high-risk]*", or "*What's the reason? Is it something in his medical history, social environment, fitness level, etc.?*"

For question prioritization, we chose to simply rank the categories by the number of elicited questions in them. For the user requirements analysis, we conducted standard qualitative analysis by first tagging individual excerpts, and then cluster them to identify emerging themes.

3.5.2 Outcomes in the use case. Our analyses revealed four main categories of questions that should be the priorities for the XAI UX to address:

- *Why*: All participants asked *Why* questions to understand the reasons behind a prediction, or risk factors of a patient.

- *How to be that:* 6 out of 9 participants asked how or what they could do to reduce the risk or prevent adverse events. This type of question was sometimes asked in a more implicit form. For example P1, as we showed above, asked about what was successfully done to other similar patients.
- *Data*: 5 out of 9 participants asked about the training data. This was somewhat surprising given that data was not an explicitly mentioned element when defining the AI tasks. It suggests that healthcare providers have a general understanding of ML models and are concerned that misalignment between training data and their own patients could hamper the system's reliability.
- *Performance*: 4 participants asked about the AI's performance. Interestingly, none of the questions mentioned standard performance metrics, but are concerned about the AI's limitations, for example, "*on what patient population might it work worse*", or "*how well does it work on non-obvious cases or patients with certain conditions?*"

Our analysis on the content about participants' intentions for asking the questions and expectation for the answer yielded the following five user requirements for the XAI UX. For a user, the XAI UX should:

- *UR1*: Help me discover (otherwise non-obvious) information to better understand patients
- *UR2*: Help me determine the most actionable next steps for the management of patients
- *UR3*: Help me decide whether I should spend more time on further evaluation or management of a patient
- *UR4*: Allow me to assess the reliability of a particular AI prediction
- *UR5*: Increase my confidence to use the AI system

Again, we emphasize that it was not a one-to-one mapping between question types and user requirements. For example, when asking a *Why* question, participants' reasons ranged between *UR1*, *UR2*, *UR3* and *UR4*. In the following steps, we will evaluate the design holistically with the set of user requirements.

3.5.3 Supplementary guidance. In different use cases, question analysis may reveal categories that are not presented in XAI Question Bank. It is worth noting that XAI Question Bank was developed with use cases mainly in the area of AI decision-support for business users.[39]. So new question categories are especially likely to appear for AI systems that are out of the scope of performing a standard decision-support task by providing predictions or recommendations. For example, in the crime analytics use case where we tested the design process, users were not only receiving AI predictions but also configuring their own model. New categories of questions such as regarding *data requirements* for the model to handle, and the *configuration process* appeared. Understanding the intended scope of the AI system early on could help preparing for the Question Analysis step.

For the prioritization task, a group could also vote on the question categories to focus based on both the insights gained from the user research and other criteria the team share. A more formal approach to actively engage the team to form a shared vision is to vote on a prioritization matrix, where one axis is about user value and the other is about feasibility, and the goal is to identify question categories that are both important for the UX and relatively easy to deliver. To complete this group voting, however, may require the team to have a level of XAI expertise to be able to effectively assess the feasibility.

3.6 Step 3: Mapping Questions to Modeling Solutions

Step 3 is for designers and AI engineers to take the prioritized question categories and start identifying an initial set of XAI techniques as functional elements that the design should include. While how a team choose to accomplish the task is flexible, we provide a mapping guide in Figure 3 to support the designer-engineer collaboration. For each question category, the guide suggests examples of explanations to answer the question, grounded in available XAI techniques, to

help designers form an understanding of what should be done and what is possible to explain (**R2**). We also link to popular code libraries for suggested XAI techniques, so that AI engineers not only have a set of concrete techniques to consider but also access to technical details. This kind of mapping artifact could serve as a boundary object to ground the discussions and consensus building between AI engineers and designers (**R3**).

3.6.1 Activities in the use case. A group of designers and AI engineers held multiple meetings to discuss the findings from the first two steps and identify candidate XAI technical solutions. The four priority question categories were used to guide the search, with an early version of the mapping guide in Figure 3. The identified user requirements also provided guidance for several decision points. Grounded in the user questions and the suggested mapping, on one side, AI engineers were able to consider the characteristics of the underlying model, on the other side, designers were able to provide input based on a deeper understanding about the users and usage contexts, both of which helped narrowing down the choices of XAI solutions.

3.6.2 Outcomes in the use case. Figure 4 is a version of the design we created in Step 4, with XAI features marked with corresponding question categories. Step 3 focused on discussions to identify functional elements without yet creating the actual design. However, for the easiness of reading, we will refer to Figure 4 for illustration purpose when we discuss specific XAI techniques.

For *Why* questions, based on Figure 3, the team quickly converged on the idea of explaining by "describing what key features of the particular instance determine the model's prediction on it" (output would look like part 3 in Figure 4) because participants were explicitly interested in patients' "risk factors". While Figure 3 suggested code libraries for state-of-the-art post-hoc XAI techniques including LIME [46] and SHAP [42], the team opted for a Feature Importance Analysis approach that the team could re-use code from a previous project.

For *How to be that* questions, the team chose to use an algorithm called Contrastive Explanation Method (CEM) [22]. As suggested in the mapping guide, CEM identifies features that if changed (increased, decreased, absent or present) could lead to the highest changes in the prediction. CEM not only identifies the top features to change, but also quantifies each changed feature's impact. While the top features to change could help satisfy *UR2: Help me determine the most actionable next steps for the management of patients*, the quantification could potentially guide users' actions to satisfy *UR3: Help me decide whether I should spend more time on further evaluation or management of a patient*–if the impact is small, it may not be worth engaging with interventions. The direct output would look like part 5 in Figure 4.

For *Data* questions, the answers are usually considered factual explanations about the training data of the model. The team carefully examined what were being asked in elicited questions and consulted the mapping guide, and decided to present facts about the data source, size of the data, distributions of key demographics, and the limitations of the data, specifically what common datasets for this kind of system are not included, as illustrated in the pop-up panel in part 1 of Figure 4.

For *Performance* questions, as discussed we noticed that the elicit questions were mostly regarding the limitations of the AI, i.e. in what situations its predictions are less reliable. The AI engineers recognized while it is possible to generate high-level descriptions about these limitations, their generalizability and utility might be questionable. The team converged on the idea of presenting the model's uncertainty level together with each prediction, so users could be alerted for cases that they should not rely on the AI's prediction.

In summary, the team reached a consensus to present in the interface feature importance information (*Why*), Contrastive explanation as generated by CEM (*How to be that*), multiple facts about the training data and confidence

information. At this point, designers were ready to create an initial design including those elements, and AI engineers could start exploring the implementation in Step 4.

3.6.3 Supplementary guidance. We acknowledge that the team had a comparatively high expertise on XAI, with the involvement of several XAI researchers. It is possible in teams where the AI engineers were less informed on the topic the collaborative problem-solving and the understanding on the mapping guide would be more challenging. A team could benefit from additional training or educational materials to understand the technical landscape of XAI, and the strengths and limitations of different XAI algorithms. In the slideshow (see supplementary material), we provide an overview of different XAI techniques and provide links to tutorials on the topic.

In this project, an RNN model was pre-selected for its performance advantage. That limited our choices of XAI techniques, for example, only post-hoc techniques like CEM can be used. For projects in an early stage, user needs for explainability could impact not only the choice of XAI technique but also the model itself, such as choosing a type of model that can generate a desired form of explanation even at the cost of performance (e.g., using a simpler, directly interpretable model).

3.7   Step 4: Iterative Design and Evaluation

In the last step, the XAI UX design is iteratively developed and evaluated with the user requirements identified in Step 2. It is important for designers and AI engineers to continue collaborating with frequent touching points (**R3**), because the translation of XAI techniques into XAI UX would often require in-parallel refinement of models or XAI techniques (**R4**). During the design iterations, it is also preferable to have target users involved to provide feedback.

This step starts with designers creating an initial version of the design that include the XAI elements identified in Step 3. Designers can refer to exemplars in the mapping guide and other references (e.g., illustrations in the linked XAI toolkits) for the chosen XAI techniques to explore the design space. Data scientists can provide additional examplers of algorithmic output to support the exploration of design.

Then the design is evaluated by enumerating on the user requirements identified in Step 2. In an early design stage, designers and the team can choose to engage in heuristic evaluation to identify gaps. In a later stage, it would be beneficial to conduct evaluative user research. One approach we found helpful is to ask participants of user research, after they experience the design prototype, to explicitly rate on each user requirements and reflect on the gaps. The gaps will then guide the next iteration of design and technical development.

It is also possible to use identified user requirements to define success metrics that guide the development of models and XAI techniques. For example, if efficiency or low cognitive workload is a primary user requirement, the team should prioritize improving the output compactness of the XAI technique.

3.7.1 Activities in the use case. After the initial design was created, it first went through iterations within the team, by reflecting on the user requirements as heuristic evaluation and discussing gaps to fill. Designers conducted evaluative research with the same group of participants and received feedback. Designers then shared the insights with AI engineers and the rest of the team to guide further improvement of the AI system. In this paper we will leave out the final system and last iteration of user research for proprietary reasons, but present an example in the first round of iteration that the team engaged in internally to illustrate the design process.
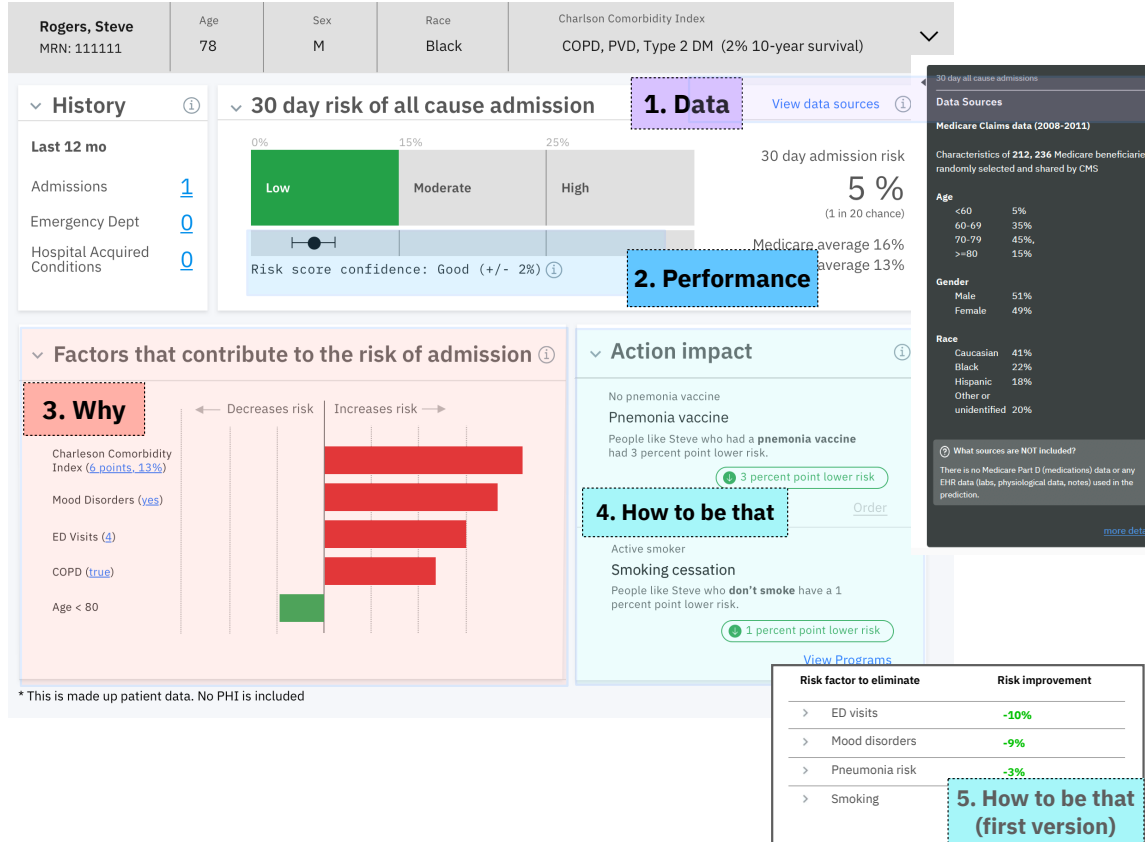
Fig. 4. Design prototype created by following the design process, with explainability features marked with corresponding user questions they answer. Part 1: *Data* explanations with a hover-over pop-up window. Part 2: *Performance* explanation with confidence information. Part 3: *Why* explanation with feature importance visualization. Part 4: *How to be that* explanation with "risk factors to eliminate" generated by Contrastive Explanation Method (CEM) then linked to actionable resources. Part 5: original *How to be that* explanation, with direct output of CEM, but lacking actionability

3.7.2 Outcomes in the use case. Figure 4 presents a design prototype of the interface for the AI system predicting healthcare adverse events, presented with fictional patient data. Different areas in Figure 4 are marked with user question types they correspond to.

One place to illustrate the outcome of an iterative design process is the contrast between part 4 and part 5, the latter of which is the first version of design to present contrastive explanations (output of CEM algorithm). It lists the top features that if changed could lower the patient's predicted risk (*How to be that*), as well as the quantity of the impact of the change on the predicted risk. However, when evaluated against *UR2: Help me determine the most actionable next steps for the management of patients*, this design was found to be critically lacking actionablity. First, the output of CEM algorithm, when suggesting top features to change (i.e. risk factors to eliminate), does not consider the practical changeability of the features or cost of changes, such as demographic factors or one's medical history (e.g., ED visit and Mood disorders as in part 5 of Figure 4). Second, the suggested risk factors only indicate "what to change", but

not "how to change". Designers could recall a quote from a participant (P4) "*I would be looking for AI's suggestion...is it plausible? Is it actionable? It needs to be associated with reasonable actions, and also with reasonable cost. I can't be like, calling everyone born in July*"

The team reflected on these limitations and came up with a new version of design (part 4 in Figure 4). The improvement first required AI engineers to update the CEM algorithm, putting constraints to prevent suggesting of a group of features that are considered unchangeable. Second, data scientists and designers worked together to identify a list of frequently appearing features in CEM output. With the input of experts in the healthcare domain, the team then created an additional layer of output by linking these features to external guidelines or resources for "how to change", whenever possible. For example, in part 4 of Figure 4, the suggested risk factor to eliminate, "pneumonia risk" was linked to information about pneumonia vaccine and link to order. "Smoking" was linked to information about smoking cessation and links to supportive programs.

For the rest of the design prototype, it answers the *Why* question with a visualization of feature importance (Part 3), which ranks features of the patient by their contribution to the patient's predicted risk, and visualizes the quantity and the direction (increase/decrease the risk) of the impact. It answers the *Data* question with a pop-up window when hovering over that presents facts about the training data, as identified in Step 3 (Part 1). Lastly, the confidence information was presented together with the the risk prediction, including both a verbal description and an error bar visualization.

3.7.3 Supplementary guidance. Most points were discussed in the beginning of Section 3.7. We call attention to the possibilities that new user questions, user requirements and the needs to add additional XAI technique can appear during evaluation, and should be incorporated into future iterations.

4 DISCUSSIONS

We first discuss some practical considerations and feedback on the adoption of this design process, then reflect on lesson learned for tackling unique challenges in designing AI UX.

4.1 Considerations and feedback for adoption

Since we started developing this work, we were confronted with the challenge of how to motivate and enable product teams with varying degrees of knowledge on the topic of XAI to adopt the design process. We decided to host workshops to bring together different roles in a product team, including designers and AI engineers, to introduce the design process in an interactive environment. A workshop would also allow us to provide richer materials for the product team to first form a shared understanding on the space of explainable AI and the possible outcomes of XAI UX. The supplementary material of this paper includes a slideshow we are using in these workshops. Following the introductory material, a team could choose to work on their own use case at the workshop. For example, a team could invite a small number of users, complete the question elicitation, work together on the question analysis and mapping to modeling solutions, then leave the workshop with a set of identified functional elements to move forward in design (Step 3). This approach would be especially useful for product teams that are interested in accelerating their development of explainability features. A team could also learn about the process through the use case we provide, then make plans to carry out the design process for their own product. Alternatively, the slideshow can be used as a reading material and followed as step-by-step instructions.

We received positive feedback from the designers and AI engineers we involved in the method development process. First, they appreciated the opportunity to *introduce a product team to the topic of XAI in a holistic way* through an overview of the technical and design spaces, concrete examples of possible design outcomes, and an actionable method to follow. One data scientist commented that "*XAI is one of those hyped topics but [what] everyone talks about is LIME and SHAP*". This comment highlights the potential "availability bias" in practitioners' choice of XAI techniques without due attention to their suitability for users. Introduction to an XAI specific design process could help product teams understand the technical space of XAI holistically and examine the choices thoroughly. The proposed method is scalable to the addition of new XAI techniques, as they could be added to the mapping guide, while the elicitation and understanding of user questions are universally applicable.

Second, designers we consulted with appreciated the *core notion of putting users' questions first.* One designer commented that even if some product teams cannot execute the design process as instructed, or they may choose to adapt it, "being able to recognize that we need to know user questions and address them goes a long way". This comment summarizes our intention well. Instead of suggesting complex guidelines on how to design different types of XAI systems, we propose to ground the solutions to when, where and how to provide AI explanations empirically in the types of user questions asked. By putting user questions in the forefront, this process centers the design and the team's shared vision for XAI UX around users instead of the AI algorithm or modeling solutions.

Lastly, many recognized the benefits of *promoting closer collaboration between designers and AI engineers*, as "a rare opportunity to bring data scientists further into the design process". Such responses resonate with prior work [25, 29, 39, 58, 59] highlighting the unique needs for AI UX design to have in-parallel development of design and modeling solutions, and challenges for designers and data scientists to collaborate due to communication barriers and a lack of shared workflows. In particular, our proposed method tackles the pain points of designer-engineer collaboration with two measures: a structured workflow with concrete touch points, and boundary objects (the question mapping guide) to bridge user needs and technical solutions. Moving forward, we hope this XAI design process provides a "petri dish" to explore how to better support designer-engineer co-creation for AI UX.

4.2 User-centered AI design: guiding the use of toolboxes of design materials

As discussed earlier, explainability is one of many AI specific design issues that are enabled and bounded by a growing space of technical solutions. In a similar vein, design issues such as fairness and uncertainty may also face the challenges of selection and translation to use a toolbox of algorithmic solutions. For example, there are a growing collection of techniques to quantify model fairness, identify areas of biases and provide de-biasing interventions [14]. The challenge of selection, to identify appropriate fairness metrics and intervention methods to use, should rely on a careful understanding on the fairness notion in a specific domain, and how the choice may impact stakeholders and the community at large. Furthermore, how to communicate output of fairness metrics to users who may be AI novice, have low numeracy, or deviate from normative views of fairness, remains a challenge of translation. For model uncertainty, both the properties of the underlying model and user characteristics should guide the selection of both the algorithm to quantify uncertainty and the method to communicate uncertainty [16].

More broadly, in current AI product practices, the choices of model and various decisions in the model development process are often dominated by technical considerations such as model performance metrics. This may pose problems down the road if the chosen model lacks means to deliver desired elements in user experiences. We believe there is a need to innovate the design processes of AI products by formulating shared design vision early on to guide the

choices of modeling solutions as well as the success metrics. To advocate for "design first" practices calls for a change of perspective by viewing the collection of AI algorithms as "a toolbox of design materials".

HCI and design researchers have discussed extensively challenges working with AI as a new and difficult design material [25, 59], including a lack of clear understanding of AI's capabilities, difficulties in envisioning uses that do not yet exist, and obstacles in prototyping. We further add to the challenges designing AI UX with a toolbox of design materials, where a collection of AI techniques are available but have different strengths, weakness and appropriate usage contexts. To form a designerly understanding of a toolbox requires not only understanding the capabilities and limitations of individual tools, but also effective ways to navigate the technical space. To tackle these challenges requires developing more actionable frameworks to organize the toolbox. For instance, there exist many taxonomies to classify XAI algorithms [11, 30] based on the technical details such as the stage of model development or type of model and data. However, they present significant disconnect from both the design goals framed as values to the users, and what is required to have a designerly understanding of the toolbox.

Our work re-frames and organizes the toolbox of XAI techniques according to the type of user questions they can answer, which could be directly linked to output of user research. This re-framing would encourage product teams, in order to navigate the toolbox of XAI algorithms, to foreground user needs and requirements instead of technical promises. More fundamentally, we believe a shift needs to happen to re-frame technical spaces of AI technologies by the human needs, values, and conditions they serve. This shift of views should start from the roots, by carefully contextualizing AI algorithms developed in academic work, articulating their strengths and weakness, and examining their potential impact for different types of users and usage contexts.

5 LIMITATIONS

The Question-Driven Design Process for XAI was developed based on feedback from 21 AI practitioners and tested with two use cases. However, we acknowledge the limitation that these practitioners worked for the same large technology company, although they were from various product lines and locations. It is possible that design practices are different in other organizations. It is also possible that this design process would not be able to account for nuances in certain AI products and can break down in certain design environments. We invite researchers and practitioners to experiment with and further refine the design process.

We are not claiming this should be the only approach to XAI UX design. Other methods have been proposed in the literature [26, 55], which offer alternative means to address the *what to explain* and *how to explain* questions. Our approach also does not offer normative guidelines on how XAI UX should be designed or evaluated. It is up to a product team to explore the design solutions and determine when the design is good enough without requiring further iterations. Future work could explore developing standard evaluation methods and metrics for XAI UX, to complement the set of user requirements identified from user research.

Lastly, as discussed, the mapping guide we provided is based on the XAI Question Bank [39]. In practices, especially for AI systems outside that scope of a standard decision-support ML system, user questions can fall outside the question bank and the mapping guide, which requires additional effort to seek solutions by collaborating with AI engineers. The guide also does not account for technical nuances. For example, there could be reliability, safety and ethical concerns to use post-hoc XAI techniques [49] in certain contexts. These are important discussions that a team should have. In short, the XAI Question Bank and mapping guide should be used as a suggestion only and a work-in-progress repository that should be continuously expanded and refined.

6 CONCLUSION

One of the critical and pervasive design issues of AI systems is their explainability–how to provide appropriate information to help users understand the AI's functions and decisions. Meanwhile, the technical field of explainable AI (XAI) has experienced a remarkable surge in recent years and produced a rich toolbox of XAI techniques. To facilitate the utilization of this toolbox, especially to *select* suitable XAI techniques and *translate* the algorithmic output into UX solutions that effectively serve user needs in real-world AI systems, we propose a Question-Driven Design Process for XAI. The method was built upon prior literature on supporting design of AI and XAI systems [25, 26, 29, 39, 58, 59], based on feedback from 21 AI practitioners, and tested and refined with two use cases. Specifically, the design process centers XAI UX design around users by grounding the user needs and requirements, design, and evaluation criteria in users' questions to understand the AI. The process supports a designerly understanding on the affordance of different XAI techniques by providing a mapping guide between prototypical user questions and exemplars of XAI techniques to answer these questions. The design process also aims to support designer-AI-engineer collaboration, as well as parallel refinement of design and technical solutions, through a shared workflow and using the mapping guide as boundary objects for consensus building. Our work offers both an actionable framework for XAI UX design, and a test-bed to address challenges in designing AI systems with a collection of algorithmic solutions as a toolbox of design materials.

REFERENCES

[1] 2017. H2O.ai Machine Learning Interpretability. https://github.com/h2oai/mli-resources.

[2] 2019. Google Cloud Model Cards. https://modelcards.withgoogle.com/about.

[3] 2019. IBM AIX 360. aix360.mybluemix.net/.

[4] 2019. Microsoft InterpretML. hhttps://github.com/interpretml/interpret.

[5] 2019. SeldonIO Alibi Explain. https://docs.seldon.io/projects/alibi/en/stable/.

[6] 2020. IBM AI FactSheets 360. https://aifs360.mybluemix.net/.

[7] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[8] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.

[9] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 275–285.

[10] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[11] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[12] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).

[13] Andrew Begel, John Tang, Sean Andrist, Michael Barnett, Tony Carbary, Piali Choudhury, Edward Cutrell, Alberto Fung, Sasa Junuzovic, Daniel McDuff, et al. 2020. Lessons Learned in Designing AI for Autistic Adults. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.

[14] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. 2018. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943* (2018).

[15] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

[16] Umang Bhatt, Yunfeng Zhang, Javier Antorán, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Gauthier Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2020. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. *arXiv preprint arXiv:2011.07586* (2020).

[17] Sander Bogers, Joep Frens, Janne Van Kollenburg, Eva Deckers, and Caroline Hummels. 2016. Connected baby bottle: A design case study towards a framework for data-enabled design. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 301–311.

[18] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 454–464.

[19] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 258–262.

[20] Tathagata Chakraborti, Sarath Sreedharan, and Subbarao Kambhampati. 2020. The emerging landscape of explainable ai planning and decision making. *arXiv preprint arXiv:2002.11697* (2020).

[21] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. 2019. Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[22] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 590–601.

[23] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.

[24] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[25] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 278–288.

[26] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing transparency design into practice. In *23rd international conference on intelligent user interfaces*. 211–223.

[27] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[28] Bhavya Ghai, Q Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL) Toward AI Explanations as Interfaces for Machine Teachers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–28.

[29] Fabien Girardin and Neal Lathia. 2017. When User Experience Designers Partner with Data Scientists.. In *AAAI Spring Symposia*.

[30] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)* 51, 5 (2018), 1–42.

[31] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, 2 (2017).

[32] Denis J Hilton. 1990. Conversational processes and causal explanation. *Psychological Bulletin* 107, 1 (1990), 65.

[33] Michael Hind. 2019. Explaining explainable AI. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 16–19.

[34] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M Drucker. 2019. Gamut: A design probe to understand how data scientists understand machine learning models. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[35] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[36] Peter Kun, Ingrid Mulder, Amalia De Götzen, and Gerd Kortuem. 2019. Creative data work in the design process. In *Proceedings of the 2019 on Creativity and Cognition*. 346–358.

[37] Vivian Lai, Han Liu, and Chenhao Tan. 2020. " Why is' Chicago'deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[38] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 2020. Co-Design and Evaluation of an Intelligent Decision Support System for Stroke Rehabilitation Assessment. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–27.

[39] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[40] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. 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*. 2119–2128.

[41] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[42] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[43] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

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

[45] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[46] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[48] Avi Rosenfeld and Ariella Richardson. 2019. Explainability in human–agent systems. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 673–705.

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

[50] Lingyun Sun, Zhibin Zhou, Wenqi Wu, Yuyang Zhang, Rui Zhang, and Wei Xiang. 2020. Developing a Toolkit for Prototyping Machine Learning-Empowered Products: The Design and Evaluation of ML-Rapid. *International Journal of Design* 14, 2 (2020), 35.

[51] William R Swartout. 1983. XPLAIN: A system for creating and explaining expert consulting programs. *Artificial intelligence* 21, 3 (1983), 285–325.

[52] Richard Tomsett, Dave Braines, Dan Harborne, Alun Preece, and Supriyo Chakraborty. 2018. Interpretable to whom? A role-based model for analyzing interpretable machine learning systems. *arXiv preprint arXiv:1806.07552* (2018).

[53] Philip van Allen. 2018. Prototyping ways of prototyping AI. *Interactions* 25, 6 (2018), 46–51.

[54] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[55] Christine T Wolf. 2019. Explainability scenarios: towards scenario-based XAI design. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 252–257.

[56] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. 2020. CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[57] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[58] Qian Yang. 2019. *Profiling Artificial Intelligence as a Material for User Experience Design.* Ph.D. Dissertation.

[59] Qian Yang, Alex Scuito, John Zimmerman, Jodi Forlizzi, and Aaron Steinfeld. 2018. Investigating how experienced UX designers effectively work with machine learning. In *Proceedings of the 2018 Designing Interactive Systems Conference*. 585–596.

[60] Bowen Yu, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, Jodi Forlizzi, and Haiyi Zhu. 2020. Keeping designers in the loop: Communicating inherent algorithmic trade-offs across multiple objectives. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1245–1257.