The Impact of Example-Based Explainable Artificial Intelligence on User Experience and Joint Activity

Thesis

Presented in Partial Fulfillment of the Requirements for the Degree Master of Science in the Graduate School of The Ohio State University

By

Maya Perlmutter

Graduate Program in Industrial and Systems Engineering

The Ohio State University

2023

Thesis Committee

Dr. Samantha Krening

Dr. Michael F. Rayo

Abstract

As machine learning gets more complex, it becomes more difficult for users to understand how the machines made their decisions. It will become increasingly important to ensure that users can understand the machine's outputs, especially in high-risk domains. Accurate systems are crucial for highly-technical domains but there is more to joint activity than accuracy; user experience is also important for ensuring that there is efficient coordination between the humans and the machines.

The purpose of this study is to examine the impact of an example-based explainable artificial intelligence (XAI) interface on trust, understanding, and performance in highly-technical populations. Approaches to increasing explainability and trust often focus on general users rather than highly-technical population in high-risk domains. This study examined the impact of showing closest matches from the training data on trust, understanding, and performance for highly-technical users. We found that providing example-based explanations significantly increased trust and understanding without decreasing performance. Adding an example-based explanation can shift participants from focusing on negative emotions to logical analysis while interacting with the ML. The ability for a human to override the decision may be a more important factor than the specific examples chosen by the XAI. Showing the most similar examples from multiple classes increased trust more than showing examples from only one class. Years

of experience was not significant for trust and understanding in the interface but was significant for trust in the human-machine team and helpfulness of the XAI cases. Our findings have greater implications for explainable artificial intelligence and trust research because it can be applied to several highly-technical domains such as the medicine, the military, and aerospace. In addition, we discuss design patterns and recommendations that can be implemented in other technical domains for future technology adoption.

Acknowledgments

I would like to acknowledge Ryan Gifford for his work on the interface design and data collection.

Vita

2020.....	B.S. Psychology, The Ohio State University
2018 – 2021.....	Undergraduate Research Assistant, Department of Psychology, The Ohio State University
2019 – 2021.....	Undergraduate Research Assistant, Battelle Center for Science, Engineering, and Public Policy, The Ohio State University
2020 – 2021.....	Undergraduate Research Assistant Center for Aviation Studies, The Ohio State University
2021 – 2023.....	Graduate Research Assistant, Department of Industrial and Systems Engineering, The Ohio State University

Fields of Study

Major Field: Industrial and Systems Engineering

Table of Contents

List of Figures

Chapter 1. Introduction

The increasing prevalence of AI is making it inevitable that we will all regularly interact with AI in our workplace and in our daily lives. Chatbots such as ChatGPT are now regularly being used by millions of users at work and at home, and other AI systems are constantly being developed by large companies such as Microsoft. While some people are hesitant to adopt AI, many have no choice but to use it because of its introduction to the workplace. People can either choose to reject it or find ways to collaborate with it, forming human-machine teams. Here, human-machine teams are teams that have at least one human and one machine that are working toward a common goal (Johnson et al., 2020; O'Neill et al., 2022). These teams should aim to have efficient team coordination, where they coordinate as a team to solve problems, make decisions, and complete various tasks (Fiore and Salas, 2004). These tasks may look different depending on the kinds of work that are being performed. However, one issue with human-machine teaming is that machines cannot be treated the same as humans. Specifically, they cannot be held accountable when they make errors, so the human teammates are responsible for the machines (Dekker, 2014). Thus, other frameworks for human-machine interaction have been proposed.

One is the idea of supertools. Shneiderman proposed that AI should be used to empower humans as super powerful tools rather than serve as teammates (2020). This

then allows the AI to extend the abilities of the user and allow the user to maintain control. In addition, the human remains responsible while maintaining authority, so responsibility-authority double binds are avoided. However, this may limit the collaboration between the human and the machine because the human maintains a high degree of control over the machine.

A broader frame for examining human-machine teaming is joint activity –any activity that is carried out by multiple parties that coordinate to work together (Klein et al., 2004). This frame is beneficial because it provides guidelines for how to maintain joint activity within a team. Maintaining joint activity involves multiple key factors and significant amounts of coordination and communication: The Basic Compact, common ground, and choreography.

First, all members that are partaking in joint activity must be aware that they are collaborating. More specifically, this is known as the Basic Compact, an agreement between all parties involved in completing a task to work toward shared goals and prevent breakdowns in coordination by maintaining common ground (Klein et al., 2004). People must be willing to relax their own goals to focus on the team's overarching goal. In addition, the Basic Compact is not an agreement that is only made at the beginning of a collaboration; rather, members must continue to reinforce it, so it is a dynamic process.

Next is common ground, which is an extension of the Basic Compact (Klein et al., 2005). Members of a team must share knowledge across the group to ensure that all are on the same page. The teams could consist of all humans or include machines. Team

members must continuously work to ensure that there is common ground to ensure that there are no breakdowns that occur.

The final dimension is the choreography of joint activity, which involves using signs, signals, actions, etc., across members and being aware of potential coordination costs that can result from breakdowns and efforts to maintain joint activity (Klein et al., 2005). An example of this could be a machine that sounds an alarm when it detects that something is wrong. Here, the machine is signaling to the human that it has detected something, and it is directing the human's attention with the sound of the alarm. Machines must be able to signal issues to humans so that human teammates can respond and direct the machines. Here, a response could be escalating a situation if there is a problem or turning off an alarm if the issue is less concerning.

However, as machine learning (ML) algorithms become more complex, it becomes more difficult for people to understand why the ML made a decision, and when or how much to rely on the ML. This can result in less trust in the system, especially in higher risk fields, because users cannot understand what the machine is doing. Furthermore, it can make maintaining joint activity more difficult because people may not understand why a machine made a signal so they may not know how to respond or how best to direct a machine. Explainable AI (XAI) aims to combat this by providing insight into how the ML made its decisions, with a goal of improving understanding, trust, and adoption of the ML system.

One form of XAI is example-based XAI in which examples from the training data are used in an explanation. Example-based XAI has been shown to help users understand

how the AI makes its classifications and where it may have limitations (Cai et al., 2019). Users can use the examples to make comparisons to what they are trying to classify, which can then support abductive reasoning by allowing them to infer the best category. However, prior research has not shown how choosing different examples from the dataset, especially whether examples should come from one vs. multiple classes, impacts trust, explainability, and performance.

Most XAI research also studies a general population of users. For example, a study on example-based explanations had users work with XAI to classify common objects such as onions or examine nutritional content of food (Cai et al., 2019; Buçinca et al., 2020). Furthermore, AI research is typically applied commercially to systems such as Amazon and Netflix, which use personalized recommendations. However, we suspect that highly-technical users have different needs for XAI systems, which impacts trust, explainability, and adoption differently than a general population of users. For example, Azari et al. found that incorporating physics knowledge on orbiting spacecraft into machine learning helped boost performance and interpretability (2021); however, it relies on the scientists' domain knowledge, so it would not be understood by general users. Consequently, the level of detail and expectations of an AI system may impact trust differently in technical populations, especially in high-risk domains.

While the accuracy of the algorithm is seen as one of the most important features of AI, performance is more than just the accuracy of the machine. People also need to have good interactions with the AI in order to have good human-machine team performance. We think high-risk tasks are likely to create different requirements for XAI

systems, which will impact usability and adoption differently than low-risk tasks. Here, we define the usability of a system as the likelihood that the system is usable, or able to be used, for activities that need to be performed (Krug, 2000). It is important to note that usability is not the same as joint activity. While usability can impact joint activity by helping ensure that a system is usable, it does not necessarily involve supporting joint activity, or coordination between members of a team.

XAI in high-risk tasks with highly-technical populations needs to be improved because there are greater risks and consequences associated with higher stakes environments when the AI makes mistakes. Designs must support joint activity architectures that are best suited for the users' context and ensure users can access their expertise when needed. Designers should also design for situations that were not anticipated so that users are supported and can recover when anomalies occur. This enables domain expertise to be integrated explicitly into technology designs, but also results in designs that enable integration of domain expertise and machine outputs at the time work is done. Errors in these contexts could risk lives, the environment, create legal and ethical issues, or result in losses of significant amounts of money.

In order to study XAI in a high-risk task with a highly-technical population, we chose a population of data analysts from an oil and gas pipeline inspection company. The task was to classify sections of pipe as healthy or not. This is a high-risk task since a misclassification could result in: 1) an environmental disaster if a damaged pipe is misclassified as healthy, or 2) the loss of millions of dollars for digging up a pipe that is misclassified as unhealthy. The participants here need XAI that is not only highly

accurate but also facilitates their understanding and decision-making process to ensure that they do not make a mistake.

The overarching goal of this work is to learn how best we can support joint activity when using XAI to improve the user experience. Specifically, we examined a group of data analysts to **determine the impact of showing the most similar examples from one vs. multiple classes has on trust, explainability, and performance in a highly-technical population on a high-risk task.** We also investigated: 1) whether the quality of the explanation impacted trust and understanding, and 2) what factors influence trust in a technical domain that uses a machine learning system.

The rest of this paper is organized as follows. First, we discuss the background research, followed by the algorithm design, study design, and procedure. We then report our results for how example-based explanations impacted trust, understanding, and performance, and we describe results from free response questions. Finally, we have an extended discussion where we describe performance outcomes, design recommendations, limitations, and future work.

Chapter 2. Background

Black Box AI

Deep learning algorithms have become increasingly popular in recent years due to successes with algorithms such as AlphaGo and their accuracy with high-dimensional data (DeepMind, 2015). However, they are often black box algorithms that provide no explanation as to how they make predictions, so users have no way of knowing whether or why the AI made an error. For example, Caruana et al. found that a neural network predicted that patients with asthma had a lower risk for pneumonia (2015). It was only after using a rule-based algorithm showing explanations that they learned the algorithm classified asthma patients as lower risk due to the aggressive treatments and shorter stay in the hospital. Without the rule-based explanation, the researchers would have had no insight into the algorithm's mistaken decisions, and patients with asthma would have been misclassified and put at greater risk. Another problematic instance is a Tesla car that identified the moon as an orange stoplight and started braking on the highway (Levin, 2021). Furthermore, Teslas, specifically in autopilot mode, have resulted in over 700 accidents since 2019, some of which were fatal (Siddiqui and Merrill, 2023).

Drivers will need to understand when and how the sensors will make errors so that they know when to override control and drive safely to prevent more accidents. Furthermore, drivers will need to be able to recover and react to unexpected errors. In all

of these cases, there is more to AI than accuracy and efficiency. While both of these components are very important, user understanding of the system is still crucial to help ensure that people not only want to adopt the AI but can also use it safely to prevent accidents.

Explainable AI

Black box algorithms are problematic because they fail to provide insight into their predictions, so users are unaware of how, why, and when errors are made. Many researchers have worked to mitigate these black boxes by introducing explainable artificial intelligence (XAI) algorithms to increase explainability.

Adadi and Berrada provided a survey on XAI methods, including various strategies for explainability (2018). Globally interpretable models provide insight into the entire model, such as Caruana et al.'s algorithm that predicted pneumonia risk (2015); whereas local interpretable models, such as Ribeiro et al.'s LIME, provide explanations for individual predictions (2016). Model-specific methods only apply to specific algorithms such as decision trees (Krishnan et al., 1999) or linear regression, while agnostic methods can be applied to any algorithm, such as rule-based explanations and example-based explanations (Caruana et al., 2015; Kim et al., 2014; Buçinca et al., 2020). This paper uses a model-agnostic method because it can be paired with different algorithms. Intrinsically explainable algorithms, such as decision trees and linear regression, enable users to quickly understand ML predictions with visual explanations. However, these algorithms are of limited use because they cannot handle complex data, and they can be unstable, so they do not scale well to real-world problems (Krishnan et

al., 1999; Sarkar et al., 2016). Explanations can take many formats, including visual representations such as saliency maps, graphs (Yang et al., 2020), or images (Buçinca et al., 2020), or they can have verbal forms such as decision-trees or rules (Krishnan et al., 1999; Caruana et al., 2015). Our work focuses on a form of post-hoc analysis known as example-based explanations, so the rest of this section will examine post-hoc analyses.

Post-hoc analyses are explanations where black box algorithms are followed by more interpretable algorithms to ensure high accuracy while increasing explainability. One common form of post-hoc analysis is including a decision tree with a prediction. For example, Krishnan et al. used a decision tree to help explain the outcomes that were provided by a neural network (2019). Similarly, Ribeiro et al. developed LIME, an algorithm that outputs a list of explanations to help users understand individual predictions, such as listing symptoms of the flu to explain a patient's diagnosis (2016).

Rather than using decision trees to explain black boxes, some researchers took a different approach: Providing examples from the training set to explain the model or the predictions. Two forms of this approach are prototype selection and example-based explanations. Prototype selection algorithms present examples that are taken from the training dataset to explain the behavior of the model (Kim et al., 2014).

Prototypes enable users to determine which features were used to make a classification because they each represent cases with similar features. Kim et al. proposed the Bayesian Case Model (BCM), which performed unsupervised clustering and generated the most representative prototypes and features, finding that it maintained accuracy while increasing interpretability (2014). Another team proposed an algorithm

called MMD-critic, which automatically presented prototypes and criticisms --prototypes that fail to represent the data well (Kim et al., 2016). The MMD-critic model differs from BCM because it provides examples that fail to represent the data alongside the best representative cases. Kim et al. also found that participants performed best in a classification task when they were presented with prototypes along with criticisms rather than prototypes alone (2016). This indicates that providing the most representative cases may not be enough, and that people may need to be able to compare and contrast good and bad examples to perform well on classifications.

Similarly, example-based XAI systems provide examples from the training data to show closest matches. Cai et al. found that participants who were shown example-based explanations felt they had a better un-derstanding of the system when asked to classify household objects such as onions (2019). Although these examples provide greater insight into the algorithm, they can also reveal limitations by presenting examples from categories that may be unrelated to a target image in a classification task (Cai et al., 2019). While Kim et al. suggest that bad examples can help users with classifications, Cai et al. claim that the bad examples only reveal limitations of the algorithm (2014; 2019). Explainability can provide users with greater awareness of the algorithm's behaviors and potentially help them understand why the algorithm makes errors, but it is essential to evaluate explainable algorithms to ensure that they are applied in the most suitable contexts.

Even so, these research teams only ex-amined whether example-based explanations were viable for black box algorithms (2019); they did not examine these

explanations in the context of trust. Buçinca et al. expanded on this research by examining the impact of XAI on trust and performance in different types of tasks, including example-based XAI, and found that people trusted different types of XAI depending on which task they completed, suggesting that context impacted which XAI was more beneficial (2020). However, they compared different types of XAI rather than comparing the same type across tasks. They also did not vary the level of explanation. Furthermore, none of these studies recruited participants in technical populations and high-stakes domains. Thus, part of our study addresses a gap in these findings by comparing levels of explanations on trust and performance in a technical population.

The Interface Design

Simply providing explanations is not enough for improved user experience and joint activity (Hoffman & Klein, 2017). Designers need to consider the users and the context in which they are using the interfaces. In higher stakes environments, the users may not have a lot of time to make decisions. The quality of information should be considered when designing interfaces; otherwise, the users may feel overloaded by the data and miss the important pieces. Some ways to prevent data overload are using positive selectivity to highlight any changes or departures from baselines (Woods, 2002). While highlighting data can add information, it can also guide the users' attention to what is most important. For example, a monitor could highlight values that have changed or are different from base values. In the case of analyzing pipes, an interface could provide

examples to help users compare and contrast healthy or problematic pipes. Another consideration could be the interjection strength of the interface.

Depending on the context, it could be helpful to have aggressive alerts that draw users' attention to it, while in other cases, a less aggressive form of alert can help users stay focused (Rayo, 2017). However, it is important to note that for all of these cases, there are still risks for obscuring data that may be important when alerts or highlights are not applied. Thus, the interfaces must be designed for the appropriate context to ensure that users are able to discern the most important pieces of information.

The Users and Usability

In addition to interface design guidelines, researchers have found that types of explanations and the needs of the users are also important considerations when evaluating AI. Hoffman et al. provided a series of success metrics to evaluate XAI, such as the goodness of the explanations, whether the user understands the AI systems, and how well the human-XAI team performs (2018). Similarly, Mohseni et al. identified three key groups of users –novices, data experts, and AI experts– and highlighted how each group had different needs and levels of understanding that the XAI designers must account for to promote successful interactions with the algorithms (2021). Here, a novice is an end-user that uses AI in their daily life has little to no expertise on machine learning systems, while a data expert has domain expertise and AI experts have a deep understanding of underlying algorithms (Mohseni et al., 2021). Both survey papers demonstrate that explainable features must be carefully evaluated to ensure that explanations are understood by users. While many of the papers on XAI have focused on novices, our

work will examine data experts that regularly interact with AI to determine whether example-based explanations are suitable for technical populations to increase trust, understanding, and performance.

Considering the types of users and the types of explanations provided also has implications for usability. Effective designs should be able to be easily used by the users they were designed for and support strong team performance. This can then impact trust, understanding, and adoption because users may be more willing to use the interfaces when they are more approachable (Kaur et al., 2022).

Trust

In XAI literature, trust has been defined in multiple ways. For example, Lee and See define trust as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability," where the agent could be a human or a machine (2004). Vereshak et al. expanded this definition, noting that trust involves three components: 1) trust requires vulnerability; 2) users need to have positive expectations for trust to develop; and 3) trust is an attitude rather than a behavior (2021). We follow these two definitions of trust because our study has participants working with agents on tasks where there is uncertainty and a need to be able to rely on the agents.

Factors that Increase Trust

One key factor in increasing trust is explainability because explanations can provide insight into the AI's decisions. Explainability can boost the predictability of an

algorithm which then increases trust (Johnson & Bradshaw, 2021). Yang et al. conducted a study on trust and XAI, where they had participants classify leaves with algorithms that provided different visual example-based explanations and found that all the explanation types increased trust but that some visual layouts such as grids were the more beneficial (2020). However, Yang et al. did not explore the impact of varying the level of explanation (e.g., examples from multiple classes). Ashoori et al. found that participants preferred transparency and human involvement (2019). Dzindolet et al. found that people were more likely to trust an agent, even after it made mistakes, if they were able to understand how the agent made its decision (2003). However, they only provided a generic explanation at the beginning of the task, and the explanations were verbal rather than visual, so their results may not generalize to other studies with XAI because explanations are typically provided throughout interactions to bolster trust. One way we address these limitations is by studying a highly-technical population in a high-risk domain, while providing examples from multiple classes.

Trust Decay

While there are studies that have examined factors that increase trust, there is also research on factors that have decreased trust. One key area of trust literature is the compliance and reliance paradigm (Meyer et al., 2014). Compliance has been defined as people responding to an alert when it is provided by automation, while reliance has been described as inaction as long as there is no alert (Rice, 2009). Both have been identified as different forms of trust in automation, but they are involved in different errors.

Parasuraman and Riley also describe decreased reliance as disuse, where users neglect or underuse the AI (1997). Multiple studies have identified two types of error that impact behavior: False alarms and misses, where false alarms are alerts given when no action needs to be taken, and misses are the absence of alerts when an action was necessary (Rice, 2009; Meyer et al., 2014; Geels-Blair et al., 2013). False alarms have been tied to compliance rates while misses have been tied to reliance, and an increase in either error will decrease rates of compliance and reliance (Goodyear et al., 2016). People are also less likely to take advice from an agent after it has made a mistake and their trust in the agent decreases more rapidly when the agent makes an obvious error (Goodyear et al., 2016; Madhavan et al., 2006). This decrease in trust may then affect the performance of the team. While the present research will not use alarm systems, we expect to see similar decreases in trust when participants see the agent make an error.

Understanding and Common Ground

XAI is successful if it enables users to understand something about the machine learning agent or how it made its decisions; otherwise, the system is not explainable, and it will not be helpful for the human. Hoffman et al. expand on these metrics by noting that the XAI must be able to help people develop good mental models of how the system works along with when, how, and why it can fail (2022). Interpretability has been used to describe understanding in machine contexts. Gilpin et al. describe the goal of interpretability as the ability to describe the inner workings of a system in a way that can be understood by humans (2018). Similarly, Doshi-Velez and Kim define interpretability

as the ability to explain or present in understandable terms to a human (2017). These goals align with the idea that simply having an explanation is not enough. The human must be able to understand those explanations; otherwise, the system is not explainable, and it will not be helpful for the human. Understanding could be linked to common ground, since humans and machines need to understand the shared knowledge, beliefs, and goals shared among them to succeed when working on tasks together (Klein et al., 2005). Understanding could also be an important prerequisite for trust, which was suggested by researchers who found that enhancing understanding in the machine resulted in higher trust (Lee and See, 2004; Chiou and Lee, 2021). Thus, it may also be important to consider understanding along with trust when investigating explainability.

Similarly, common ground is where members of a team share knowledge across the group to ensure that all are on the same page. In the case of machines, they need to continue to provide enough explanation to ensure that users have an understanding of what is going on. Team members must continuously work to ensure that there is common ground to ensure that there are no breakdowns in coordination.

The fundamental common ground breakdown is a form of a breakdown where one member leaves the group while the others still believe that the Basic Compact is still in effect among all members (Klein et al., 2005). For example, a fundamental breakdown could be an AI agent that adopts, adjusts, or relaxes its goal set in a way that breaks the Compact while failing to update its teammates about the changes. Here, AI could have an unexpected software update that impacted its usual responses. Members would only be aware of a common ground breakdown when the AI makes a major error. XAI, on the

other hand, could provide explanations that give members insight into what it is capable of doing and how it understands the current world. Users are more likely to be aware of changes to the machine because of the increased observability provided by the explanations. The explanations can then facilitate how members direct the XAI because they can understand how and why it makes decisions. This could be through examples it provides from the training data, visuals such as heat maps or graphs, or visualizations of the model itself. XAI could also be employed to show humans why there is a problem by highlighting contrasts so that the operators can determine whether to attend to the issue or ignore it. Thus, XAI could be used to facilitate signals to help maintain common ground and provide users with important information.

How common ground is maintained can depend on the context. For example, Chiou et al. examined the impact of explanation-based communication strategies on human-robot team performance in the context of a search and rescue task, and they found that robots providing relevant explanations paired with moderate levels of communication helped teams maintain common ground and have stronger performance overall (2021). This suggests that there are some levels of communication and explanations that are more effective than others. Chiou et al. note that having a robot provide explanations for every event does not necessarily support team performance, which would make sense because communication can result in time costs, especially in time-critical contexts (2017). Frequent communication may help maintain common ground, but it could also take time away from emergency responses

Performance and Informativeness

Having strong performance is crucial to human-machine teaming. The machine must be able to provide benefits for humans rather than hindering them with time and coordination costs. One framework for team performance is macrocognition: detecting, interpreting, and responding (Patterson & Hoffman, 2012). First, people need to be able to detect when something is wrong or anomalous. In the case of oil and gas pipelines, the analysts need to be able to determine when a pipeline is anomalous and needs to be examined. The machine predictions also need to facilitate detection by clearly showing users why something may be problematic. This could be through images, alarms, or a combination of signals (Patterson & Hoffman, 2012). Next, is interpreting signals. Users must be able to understand a situation by making sense of machine signals and generating hypotheses for causes. XAI could facilitate the sensemaking process by providing users with insight as to how it reached a prediction so that users can gain more information. Finally, users need to be able to appropriately respond to events in a timely manner, especially in high stakes domains. Users need to be able to determine if there is a true anomaly and how best to respond to it, or conversely, determine when they should proceed with their normal work because there was a false alarm.

One important consideration for strong performance is informativeness. Rayo et al. define informativeness as what the automation detects and interprets along with what it effectively conveys to the users (2022). In the context of machine learning, this is important to measure because users can over-rely on a machine –continue to agree with a machine's predictions even when it is wrong. Misleading predictions could also confuse

users and delay the decision-making process, which would be especially problematic in a high stakes domain. On the other hand, a machine that makes enough errors could be ignored which could be problematic when the machine detects that something is actually wrong. Explainable AI must be able to provide users with enough information so that they are aware of when the machine is making errors.

Machine-Fitness Assessment

Machine fitness assessment involves determining whether a machine's predictions or inferences about the world are well aligned to the actual state of the world (Rayo et al., 2020). Here, a world is a workplace, such as the oil and gas pipeline inspection company as examined in this work. Users need to be able to determine when a machine is well aligned to the world and when it is not. Otherwise, they risk over-relying on the machine, blindly agreeing even when the machine makes egregious errors. A machine that provides explanations that are well tailored to the context could help users be more aware of when the machine is more likely to make errors and subsequently follow their own judgments. Knowing when and when not to rely on the machine could boost team performance and mitigate issues with over- and under-reliance. However, it is important to note that there are few studies that have examined this work (Rayo et al., 2020; Rayo et al., 2022). While our work does not directly touch on machine fitness assessment, we do consider performance in the context of appropriate use compared to overreliance and under-reliance.

Designs

Design patterns are designs that can be applied to different contexts or problems. They can also be introduced to combat recurring challenges. For example, a recurring problem could be that a user in a time-critical situation has limited attention, and a design needs to help guide them to the most important cases. A design pattern that addresses attentional demands could be having a display that highlights changes and contrasts from baselines so that users can have a quicker time detecting problems and prioritizing responses accordingly. In the case of our work, the data analysts from the oil and gas pipeline inspection company need to be able to correctly identify when a pipe is healthy or problematic. The interfaces were designed to help them determine differences by providing examples. In one of our designs, we included contrasting examples so that the analysts could compare and contrast the cases to make a more informed decision. Design patterns are important to examine because most designs are tailored to specific contexts. Abstracting patterns from designs can help generalize patterns to other contexts.

Amershi et al. created a list of eighteen general design patterns for human-AI interaction that included guidelines for before and during interaction along with when a design is wrong and interactions over time (2019). These guidelines included designs such as clearly showing the system's capabilities and limits, showing contextually relevant information, providing explanations, and notifying users about changes. These guidelines align to previously discussed topics such as machine fitness assessment, XAI, and common ground. However, it is important to note that Amershi et al. only tested these guidelines in everyday products such as music recommenders, activity trackers, and

web search; none of these systems specifically applied to high stakes environments. Rayo provided design heuristics that were meant to apply across all industries. These are: supporting detection, supporting sensemaking, supporting replanning, making the Basic Compact explicit, and designing for common ground (2017). The designs must support the full perception-action cycle while reinforcing the most important facets of joint activity. In addition, the designs must help the human-machine team maintain good performance.

While our work will not address all of these guidelines, we will discuss the design patterns we focused on in the extended discussion.

Our Work

This paper aims to examine factors that increase a human's trust and understanding in explainable AI, with a focus on highly technical populations. Neural network algorithms have previously been viewed as less explainable because they include many hidden layers, so people are unable to see which layer or feature the algorithm used to make its decision (Guidotti et al., 2018). This lack of visibility into decision-making often makes it difficult for humans to trust the agents. The algorithm used in this study strives to mitigate these issues by providing examples from the training set as a form of explanation, like prototype selection and example-based explanations. Specifically, the explainable interfaces will show the user how the algorithm classifies input data based on the closest matches that it finds in the training data. This will provide the user with insight into the features that the algorithm is using to make its classifications. We expect

that the increased explainability will not only increase trust in the machine, but that it will also increase performance compared to an interface that provides no explanations.

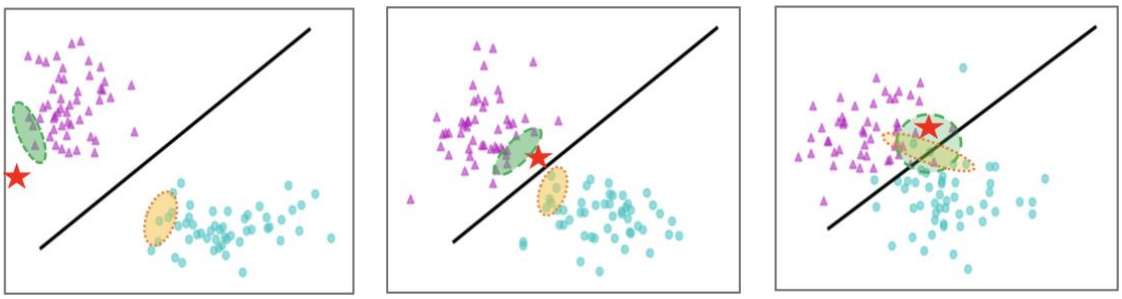
Chapter 3. Methods, Results, and Discussion

Methods

Choosing One Example from One vs. Multiple Classes

Participants were asked to classify oil pipeline images as normal ("nothing") or abnormal ("anomaly"). They were provided with predictions made using neural network models and were asked whether they agreed with the predictions.

Example-based XAI research has not thoroughly determined which examples should be shown to the user along with the impact the chosen examples can have on the user's trust in the ML system. To explore this, we created an XAI system that could show participants the most similar examples from the dataset from multiple classes.



(a) (b) (c)

Figure 1. Choosing examples for XAI.

Consider Figure 1 above. The scatter plots show data in two dimensions. There are two classes: magenta triangles and cyan circles. Assume that the red star was not in the training data and needs to be classified. The ML classifies the red star as belonging to the blue circle class. An XAI system that shows users the closest training data from the magenta class would show all magenta data within the green dashed ovals. An XAI system that shows the users the closest training data from the cyan class would show all cyan data within the orange dotted ovals.

Participants and Design

We conducted a within-subjects study in which we investigated the effect of three interfaces on the human's experience with the agent. Our 24 participants were data analysts from a pipeline inspection company with varying levels of experience at the company. We divided experience level into four groups based on the amount of time they worked as a data analyst at the company: 2 participants in the 6 months - 2 years group; 13 participants in the 2 - 5 years group; 6 participants in the 5 - 10 years group; and 3 participants in the 10+ years group.

Participants were asked to classify oil pipeline images as normal ("nothing") or abnormal ("anomaly"). They were provided with predictions made by a neural network model and were asked whether they agreed with the predictions.

To predict whether an image of a pipeline section was considered normal or an anomaly, a simple Convolutional Neural Network (CNN) was used. Images of 100x100 were passed through two convolutions, each followed by a batch normalization layer and

a max pooling operation that down sampled the feature space by a factor of 2. The image features were then flattened and fed into 4 fully connected layers with hidden node sizes of 50, 25, 6 and 2. The last layer used a Softmax activation to output the probability of an image being normal and an anomaly.

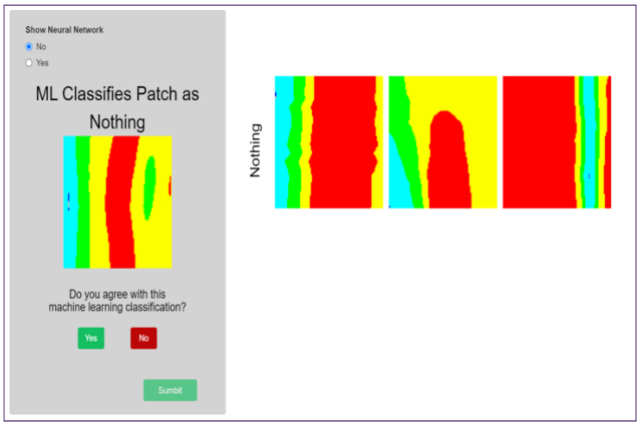
Participants interacted with three interfaces that had varying levels of explanation to classify the pipeline images. One interface was a baseline that provided no explanation, another interface showed the three closest examples from the class that it selected as its prediction, and the final interface provided users with the three closest classifications from both the 'nothing' and 'anomaly' classes. We describe the interfaces below.

A.

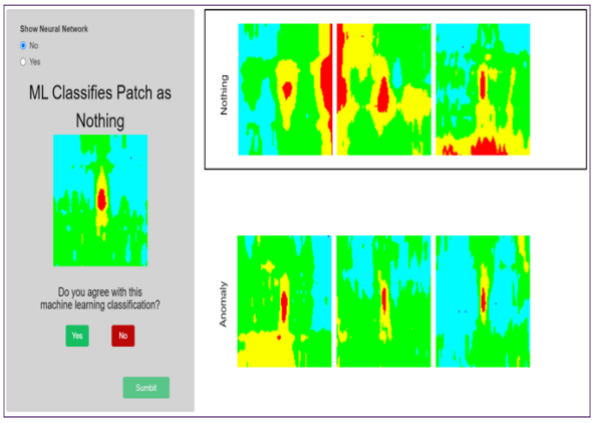
Continued

Figure 2. A continued figure. Interfaces. Part A. Baseline interface; Part B. XAI showing closest matches within one category (Interface XAI-1); Part C. XAI showing closest matches in both categories (Interface XAI-2).

Figure 2 continued



B.



C.

Interface Baseline: The participant only saw the current input image and the suggested classification by the model. The interface did not provide any further explanation. See Figure 2A.

Interface XAI-1: The XAI showed closest matches within one class. For each case, the ML classified an image. The participants saw the current input image, the

suggested classification by the algorithm, along with the closest matching images from the suggested class in the training data. For example, if the algorithm predicts that an input image is an anomaly, then the explanation would show three of the closest cases of anomalies from the training data. This is like showing the yellow oval examples to a user from Figure 1. See Figure 2B for the interface layout.

Interface XAI-2: The XAI showed closest matches in both classes. The participants saw the current input image, the suggested classification by the algorithm, along with three examples of the closest matching images from the training data for both classes. This is like showing both the yellow and purple oval examples to a user from Figure 1. See Figure 2C for the interface layout.

Procedure

The procedure followed by all participants is provided below. All participants experienced the interface variations in a randomized order.

1. Consent.
2. For each of the 3 interfaces:
    a. For each of 4 cases (e.g. trials):
        i. Participants were shown the local image of the pipeline and the ML agent's classification. They were asked whether they agreed with the classification.
    b. Questionnaire about their experience with the interface.
3. Questionnaire comparing their experiences with the interfaces.

Human Experience Measures

After interacting with each interface, participants completed a questionnaire (Step 2b) about: 1) the extent to which they agreed with the classifications; 2) the extent to which they trusted that algorithm; 3) the extent to which they trusted the human-machine team; 4) whether viewing similar cases helped them understand the agent's classification; and 5) whether they understood why the agent made the classification that it did. These Human Experience measures were measured continuously on a scale of [0:10]. Values of 10 indicated strong agreement, strong trust, helped a lot, and strong understanding. Values of 0 indicated strong disagreement, strong distrust, no help at all, and no understanding.

After interacting with each interface and at the end of the study (Steps 2b and 3), participants ranked the interfaces by trustworthiness and understanding. Participants were also given the option to explain factors that impacted their experience working with the interfaces. For example, participants were prompted, "Please provide comments on what impacted your trust in this interface." Participants were not primed or given any information for these written responses.

Objective Measures

While participants interacted with the interfaces, objective performance metrics were logged to data files. Whether the participant agreed with or overrode the ML's classification, along with whether the resulting decision was accurate, were recorded.

Results and Discussion

The results will begin with the Human Experience data from Step 2b and 3, followed by the Objective data collected from Step 2ai. At the end, we discuss free response answers from Steps 2b and 3.

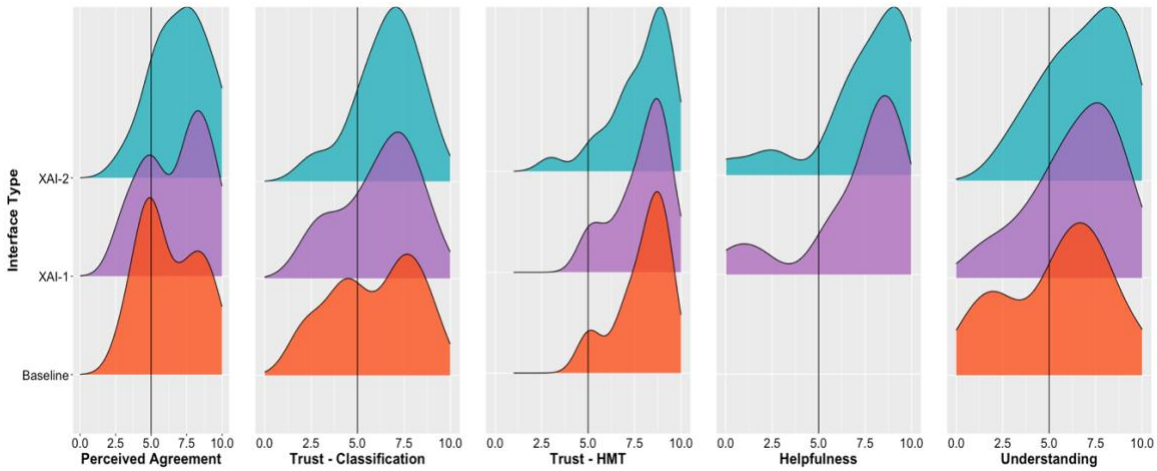


Figure 3. Distribution of human experience measures (Step 2b).

Figure 3 shows the distributions of the Human Experience measures (Step 2b). A line is drawn at the middling value of 5 for each of the factors. All of the Baseline measures and many of the XAI-1 measures resulted in bi-modal distributions, whereas XAI-2 were mostly unimodal. This indicates that the Baseline and XAI-1 interfaces caused groups of participants to react quite differently – with one set of participants having a worse user experience. However, the XAI-2 interface had a more uniform user experience and did not provoke a more negative reaction from some participants. Overall,

participants understood the XAI-2 interface the most, indicating that it aids understanding to show examples from multiple classes. Trust in the human-machine team was much higher across all interfaces that trust in the machine alone. Most users found interfaces that showed similar examples to be helpful in understanding the ML, while a few participants found the examples to provide very little help. The perceived agreement of the Baseline and XAI-1 interfaces were bi-modal and almost the reverse of each other; whereas the XAI-2 interface was unimodal and more positive on average. It is important to note that participants were not asked about helpfulness of the similar examples when interacting with the Baseline interface since similar examples were only shown in the XAI interfaces.

Trust

While participants consistently ranked the XAI-2 interface as the most trusted at the end of the study, our results indicate that the ability for the human to override the ML's decision is more important to overall trust than the interface or years of experience.

Trust Rankings

At the end of the study in Step 3, participants were prompted: Please rank each interface you trusted in order of most to least. We conducted a Kruskal Wallis test and found a significant difference in ranking between the interfaces ($\chi^2(2) = 11.39$, $p = 0.003$). Participants consistently ranked the XAI-2 interface as the most trustworthy, followed by XAI-1, followed by the Baseline. See Figure 4.

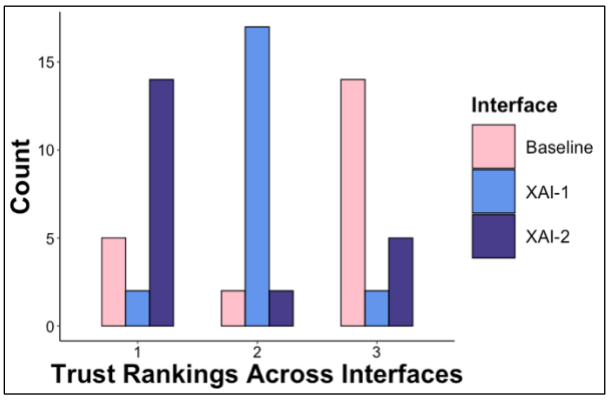


Figure 4. A grouped bar plot comparing rankings across the interfaces. A rank of 1 indicates that an interface was considered the most trustworthy. (Step 3)

Trust in the Classification

Immediately after interacting with an interface (Step 2b), participants were asked: How much do you trust the machine learning classification? Given the bi-modal nature of the Baseline distribution (Figure 3), we were surprised the assumptions for an ANOVA were met. We did not find significant differences in trust in the ML's classifications by interface from the Step 2b data, even though they ranked the XAI-2 interface significantly higher in trust at the end of the study ($F(2, 69) = 0.719$, $p = 0.491$) (Step 3). It is possible that the effect size was smaller for the Step 2b data since it did not directly compare interfaces. Another possibility is that participants viewed the ML and XAI interfaces as different entities, which garnered different levels of trust.

Also, a Kruskal-Wallis test found that there were no significant differences in trust in the ML's classification by years of experience (Step 2b, $\chi^2(3) = 3.43$, $p = 0.330$).

All participants had a moderate- to high-level of trust in the machine regardless of their level of experience.

Our next step was to determine which features impacted trust in the classification, especially since interface was not significant in the Step 2b data. For each interface, we trained a Random Forest Regression model to predict trust in the ML's classifications using years of experience, objective performance from Step 2ai, Step 2b data (including perceived agreement and understanding), and Step 3 data. The r^2 value for each model was: Baseline $r^2 = 0.95$; XAI-1 $r^2 = 0.95$; XAI-2 $r^2 = 0.90$. The **most important feature for predicting trust** in the ML's classifications for all interfaces was the participant's **perceived agreement** with the classification. For the XAI-2 model, the participant's nervousness against robotics was also an important factor in trust, although much less than perceived agreement (i.e., *I feel that if I depend on robots of artificial intelligence too much, something bad might happen*.)

Trust in the Human-Machine Team

In Step 2b, participants were asked: How much do you trust the human-machine team (which includes you)? Interestingly, there was not a significant difference between interfaces when examining trust in the human-machine team. However, we found that people had higher trust in the human-machine team than in the machine learning agent alone. There was a significant and positive correlation between trust in the machine and trust in the human-machine team across all interfaces ($r(70) = 0.414$, $p < 0.001$). Additionally, we conducted a Wilcoxon Signed Ranks test and found that trust was rated

significantly higher for the human-machine team compared to the machine's classifications ($Z = -6.186$, $p < 0.001$). These findings indicate that **the ability for a human or override or contribute to the decision is a more important factor for trust than the examples shown** in an interface's explanation. See Figure 5.

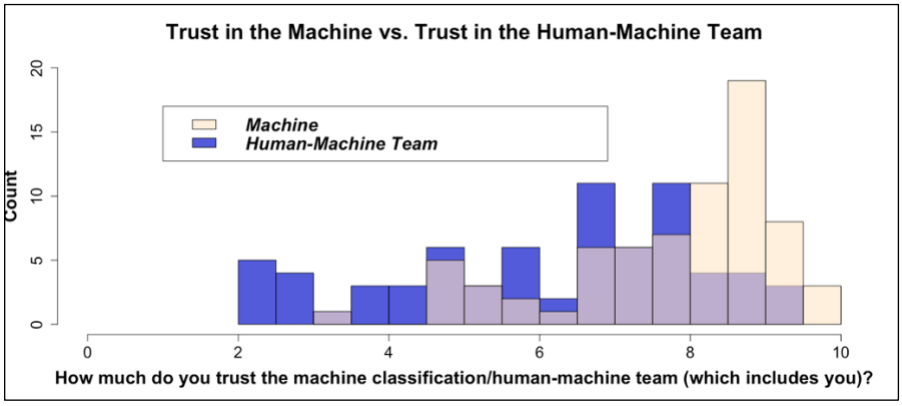


Figure 5. A comparison of trust ratings for the machine classification and the human-machine team. People trusted the human-machine team more than the machine's predictions alone. (Step 2b)

Users in technical domains may have a greater need for control over the ML because they work on high stakes problems. In this case, the personnel at the oil and gas pipeline inspection company could risk environmental disaster with pipe leaks, so they may want to be more involved in the decision-making process rather than rely on an ML model.

A Kruskal-Wallis test found that there was a significant difference in trust in the human-machine team by years of experience ($\chi^2(3) = 15.68$, $p = 0.001$). Trust in the

human-machine team decreased as the years of experience increased. The 6 month – 2 years group was the most likely to trust the team, followed by the 2 – 10 years groups, and the 10+ years group which had the lowest trust. This is interesting because years of experience was not significant for trust in the ML – it only becomes significant when a human is added to the team. This suggests that those with more experience may not find as much solace in adding a human to the decision-making process compared to less experienced personnel.

Next, we determined which features impacted the participants' trust in the human-machine team. For each interface, we trained a Random Forest Regression model to predict understanding in the ML's classifications using features including: years of experience, objective performance from Step 2ai, Step 2b data (including perceived agreement and understanding), and Step 3 data. The r^2 value for each model was: Baseline $r^2 = 0.87$; XAI-1 $r^2 = 0.90$; XAI-2 $r^2 = 0.90$. The Baseline and XAI models did not share the same feature importance; this suggests that adding an example-based explanation changes how people trust and think about an ML agent. For the Baseline interface, the most important feature was a measure of the participant's **nervousness** against robotics (i.e., *I feel that if I depend on robots of artificial intelligence too much, something bad might happen.*) However, for the XAI interfaces, the most important feature in predicting trust in the human-machine team was the extent to which the participant **understood** the ML's classification. This suggests that adding **example-based explanations shifts participants from focusing on emotions to logical analysis** while interacting with the ML agent. As people interact with ML more and more in their

daily lives, it will become necessary to identify means of reducing a human's stress, nervousness, and anxiety. In terms of rhetoric, this can be thought of as a shift from pathos to logos.

Understanding

We found that levels of understanding differed across interfaces and years of experience, with higher levels of understanding being associated with interfaces that provided explanations and less experienced groups.

Understanding Rankings

In Step 3, participants were asked to: Please rank each interface you understood in order of most to least. We conducted a Kruskal Wallis test and found significant differences in ranking between the interfaces ($\chi^2(2) = 9.42$, $p = 0.009$). Participants ranked the XAI-2 interface as the most easily understood, followed by XAI-1, and lastly the Baseline. **Given the same underlying ML agent, people felt they understood the agent the most when it provided examples from multiple classes**. See Figure 6.

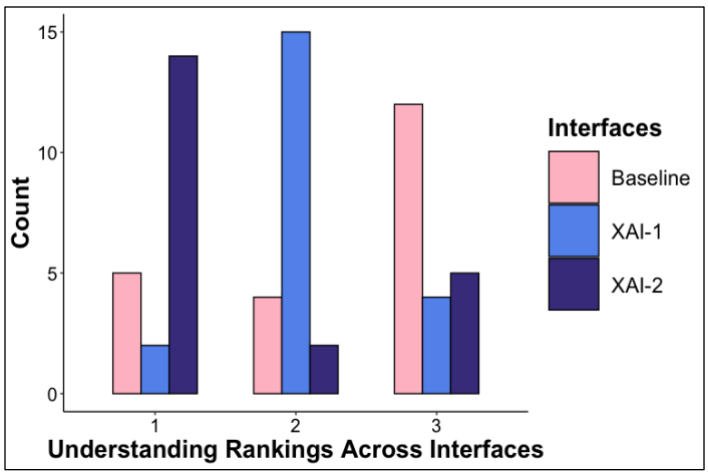


Figure 6. A grouped bar plot comparing rankings across the interfaces. A rank of 1 indicates that an interface was considered the most easily understood. (Step 3)

Understanding the Machine's Classifications

In Step 2b, participants were asked the following question: To what extent do you feel you understand why the ML made the classification it did? Overall, participants reported a moderate level of understanding. An ANOVA reported weak differences between interfaces ($F(2, 69) = 2.520$, $p = 0.088$).

A Kruskal-Wallis test found there were significant differences in understanding across years of experience ($\chi^2(3) = 18.908$, $p < 0.001$). The 2 – 5 years group and the 5 – 10 years group had the highest ratings, followed by the 6 month – 2 years group and 10+ years group. Participants with middling experience reported more understanding than the most and least experienced analysts.

We determined which features impacted the extent participants understood the ML agent's classifications. For each interface, we trained a Random Forest Regression

model to predict understanding in the ML's classifications using features including: years of experience, objective performance from Step 2ai, Step 2b data (including perceived agreement and understanding), and Step 3 data. The r^2 value for each model was: Baseline $r^2 = 0.88$; XAI-1 $r^2 = 0.92$; XAI-2 $r^2 = 0.90$. The Baseline and XAI models did not share the same feature importance. For the Baseline interface, the most important features for predicting understanding were trust in the classification and perceived agreement. For the XAI interfaces, the **helpfulness** of showing similar examples and **trust in the human-machine team** were the **most important features in predicting understanding**.

Correlations Between Trust and Understanding

We examined the Spearman Rho's correlation between the rankings for trust and understanding (Step 3). We found that there was a perfect correlation between trust and understanding for the XAI-2 interface ($r(19) = 1.00$, $p < 0.001$), and strong positive correlations for the Baseline ($r(19) = 0.904$, $p < 0.001$) and XAI-1 ($r(19) = 0.818$, $p < 0.001$) interfaces.

We also found positive correlations between trust and understanding when participants were evaluating individual interfaces in Step 2b. We found there were moderately positive correlations between trust in the machine and understanding ($r(70) = 0.348$, $p = 0.003$) and trust in the human-machine team and understanding ($r(70) = 0.482$, $p < 0.001$). These findings suggest that providing explanations from multiple classes positively impacts both trust and understanding, and that the two are strongly related.

This also can imply that highly-technical **users trust the interface more when they think that they understand it**, and vice versa.

Perceived Agreement

In Step 2b, we asked participants the following question: *How much do you agree with the ML agent's classification?* It is important to note this is perceived agreement rather than objective agreement (Step 2ai). Overall, participants reported a moderately high level of agreement with the machine ($M = 6.72$). There were no significant differences in perceived agreement across the interfaces ($F(2, 69) = 0.832$, $p = 0.440$). The years of experience were also not significant for perceived agreement across interfaces ($\chi^2(3) = 4.59$, $p = 0.204$).

We trained a Random Forest Regression model to predict perceived agreement using years of experience, objective performance from Step 2ai, Step 2b data (including trust and understanding), and Step 3 data. The r^2 value for each model was: Baseline $r^2 = 0.96$; XAI-1 $r^2 = 0.97$; XAI-2 $r^2 = 0.92$. The **most important feature for predicting perceived agreement** for all interfaces was the participant's **trust in the classification.**

Helpfulness

In Step 2b, for both XAI interfaces we asked participants the following question: *To what extent did showing similar cases help you determine the input's classification, regardless of whether you agreed with the ML classification?* Overall, participants had

high perceptions of the machine's helpfulness. The two XAI interfaces were not significantly different in helpfulness ($\chi^2(1) = 0.207$, $p = 0.65$).

There were significant differences in helpfulness based on the participants' years of experience ($\chi^2(3) = 10.67$, $p = 0.014$). Specifically, the 2 – 5 years group and the 5 – 10 years groups had significantly higher ratings than the others, suggesting that the interfaces were more helpful for personnel with moderate levels of experience.

We trained a Random Forest Regression model to predict the helpfulness of showing similar examples in the interface using features including: years of experience, objective performance from Step 2ai, Step 2b data (including trust and understanding), and Step 3 data. The r^2 value for each model was: XAI-1 $r^2 = 0.87$; XAI-2 $r^2 = 0.83$. The **most important feature for predicting helpfulness** for both XAI models was the extent that participants **understood** why the ML agent made each classification.

Objective Performance

Performance was divided into four categories:

1. The human agreed with the machine and the machine was correct.
2. The human agreed with the machine, but the machine was incorrect.
3. The human disagreed with the machine, but the machine was correct.
4. The human disagreed with the machine, and the machine was incorrect.

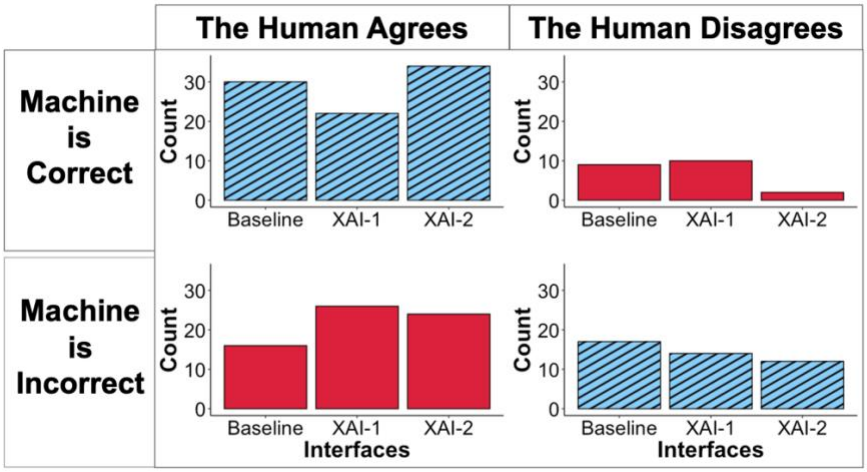


Figure 7. Human-machine team performance across interfaces.

Any response that falls within the first or fourth categories (the human agreed with the machine and the machine was correct; the human disagreed with the machine, and the machine was incorrect) would indicate a successful human-machine team (see blue lined bars in Figure 7). Although the fourth category has the human disagreeing with the machine, the goal is to make a correct decision, regardless of agreement. Thus, if the human overrides a machine's prediction and makes the correct decision, the team is successful. Note: This is objective agreement recorded from Step 2ai, not perceived agreement from Step 2b.

Participants agreed with the machine 70.4% of the time (categories 1+2), regardless of whether the machine's prediction was correct, with an overall success of 59.7%. Figure 7 summarizes the performance distributions below.

The four categories of objective performance were not significantly different across the interfaces ($\chi^2(3, N = 216) = 11.46, p = 0.075$). This means that **adding in an example-based XAI interface did not reduce objective performance**.

The four categories of objective performance were significantly different depending on whether the case was of an anomalous or healthy portion of pipe ($\chi^2(3, N = 216) = 15.83, p = 0.001$). Regardless of case type, the human and ML were both correct (Category 1) approximately 40% of the time. It was approximately 10% more likely for analysts to agree with the machine when it was wrong for non-anomalous cases compared to anomalous cases (non-anomalous ~35%; anomalous ~26%). This may point to a risk estimate by analysts who would rather agree with the ML when it says there is an anomaly, even if the case is truly non-anomalous. It was also approximately 10% more likely for analysts to disagree with the ML when it was correct for non-anomalous cases (non-anomalous ~14%; anomalous ~4%). This may also point to a risk preference by analysts who would rather classify a pipeline section as anomalous (unhealthy) rather than risk an environmental disaster. It was approximately 18% more likely for analysts to disagree with the ML when it was wrong for anomalous cases (non-anomalous ~12%; anomalous ~30%). This again points to the idea that analysts would rather err on the side of an anomalous classification. **In high-risk domains in which the the consequences of misclassification differ by class, highly-technical analysts are likely to favor the high-risk classification, regardless of if that means agreeing or disagreeing with the ML**.

We expected people to treat classifications differently because the risk associated with each class is different. This expands beyond the oil pipeline domain into many safety-of-life applications such as medicine, e.g., the risks of diagnosing a patient with a disease or not. Misclassifying a non-anomalous pipeline section can lead to a large monetary loss, while a misclassification of an anomalous pipeline section can lead to an environmental disaster. Anomalies also occur infrequently, so personnel are more likely to be more cautious when classifying cases as anomalies.

When looking closer at whether a specific case presented to a participant was an anomaly, a difference did show up. The four categories of objective performance were significantly different across the interfaces for non-anomalous cases ($\chi^2(6, N = 119) = 14.19, p = 0.028$), but not for anomalous cases ($\chi^2(6, N = 97) = 2.74, p = 0.841$).This again suggests that analysts did not treat the two classifications equally.

The four categories of objective performance were not significantly different across the analysts' years of experience ($\chi^2(9, N = 216) = 6.80, p = 0.658$). This suggests that performance was roughly equal regardless of how much experience an analyst had. The main difference across groups was their user experience.

We trained a Random Forest Classification model to predict objective performance using years of experience, objective agreement from Step 2ai, Step 2b data, and Step 3 data. The accuracy was 0.85. The most important feature for predicting objective performance was **objective agreement**.

Objective Agreement

Accuracy is defined as whether the human-machine team ended up with the correct answer, even if the human had to override the ML's classification; *agreement* is whether the human agreed with vs. overrode the ML's classification.

Objective agreement was not significantly different for each interface ($\chi^2(6, N = 216) = 5.51, p = 0.064$). People were more likely to agree with the ML, regardless of whether it was correct (70.4% agreement; 29.6% disagreement) (Figure 9). Adding XAI to the interface did not change or harm whether people agreed with the ML.

Objective agreement was significantly different for whether the ML was correct or not ($\chi^2(1, N = 216) = 10.18, p = 0.001$). People agreed with the ML more often when the ML was correct (~57%) and disagreed more when the ML was incorrect (~67%). Objective agreement was not significantly different across the analysts' years of experience ($\chi^2(3, N = 216) = 3.42, p = 0.33$). Less experienced analysts did not over-rely on the ML to compensate for a lack of experience compared to experienced analysts. We trained a Random Forest Classification model to predict objective agreement using years of experience, objective performance from Step 2ai, Step 2b data, and Step 3 data. The accuracy was 0.995. The most important feature for predicting objective agreement was case type (i.e., anomalous vs non-anomalous).

Interestingly, while there is a significant association between perceived and objective agreement ($r(214) = 0.301, p < 0.001$), it is a weak, positive correlation. This indicates that people do not accurately estimate how often they agree or disagree with the ML.

Free Response

During Step 2b, participants were given the option to write free-response answers regarding their interaction with the agent. For understanding, participants were prompted, "Please provide comments on what impacted your understanding of this interface." For trust, participants were prompted, "Please provide comments on what impacted your trust in this interface."

Two researchers analyzed the responses by identifying key themes from the written answers. Researchers tallied which and how many responses fit within each theme and verified agreement. Free response questions were optional; thus, not all questions received the same number of responses. We identified the following themes: 1) priming; 2) training and education; 3) level of explanation; 4) interface presentation; and 5) machine errors.

Priming

When studying highly-technical populations at work, we have found priming to be especially important. Designers should verify that participants know the study's goal is not to replace or change their workflow, tools, or job. Two participants were concerned the ML used in the study would be integrated into their workflow without being sufficiently tested. Previous research has found that people resist new technology because they fear it will result in change when they already worked hard to adjust (Juma, 2016).

One participant claimed that trust was irrelevant because they would never use ML in their job no matter what. While this was only one participant, it is important to

note that there are people with anti-ML sentiments, even among the most technical populations. This could stem from a fear of losing their job, a loss of control due to the ML making decisions, or even responsibility-authority double binds in which people could be held responsible for machine errors.

Training and Education

Eight responses mentioned that increasing the number of cases they saw while being trained on the tool would boost trust (5 participants) or understanding (3 participants). With more examples, trust may decay at a slower rate.

Three participants had unrealistic expectations of ML, where they would "fully" trust the machine if it were 100% accurate, but not until then. Having clear explanations about how the machine will never be perfect could help mitigate these expectations and shift trust to a spectrum rather than a binary decision between full and no trust. Future systems integration processes should ensure that users are familiarized with new systems through training to ensure that all users understand the system's capabilities.

Level of Explanation

Although multiple participants noted that more interaction time with each interface would boost trust and understanding, they still reported having more trust and understanding –even with limited interaction time– when explanations were provided.

Five participants mentioned that having the XAI improved either trust (1) or understanding (4). Conversely, a participant reported that having no explanations decreased their understanding.

> "Having the examples certainly was the best thing, obviously the more you can rely on patterns and previous experience, the better you can make a call. It's the same in regular analysis work, if you have historical data that can give you additional info it can only benefit you in the long run, even if it may be conflicting at first."

> "… it['s] very important having some examples for final classification, that could increase my trust in AI."

> "Seeing what it is using as a point of reference helps in understand[ing] why it['s] classified as such."

> "[T]here really wasn't anything to go on [with the Baseline interface], I was just looking at signals one at a time with no real understanding [of] why it did what it did. All I knew is that it didn't get them right, so I had no reason to believe there weren't other errors."

Example-based explanations increased trust and understanding, while the lack of XAI decreased understanding. This is important for the adoption of ML.

Interface Presentation

Regardless of other factors such as the amount of explanation or use of human input, we found that the interface presentation impacts trust and understanding. Specifically, many participants noted that the XAI interfaces differed from the appearance of the user interface they currently use in their job, and that the terminology was different. Three participants reported that these differences impacted their trust while five said it impacted their understanding. Future designs should have interfaces be tailored to the workplace to ensure that designs are not significantly different from existing interfaces. This can then help with a smoother transition between the capabilities when they are introduced to the workplace. Again, having training and education on the new systems can also help clear confusion and ensure that users understand how the new designs work.

Machine Errors

Four participants reported that their trust decreased after seeing the machine make an error.

> "[I]t did miss a pretty obvious signal and that will make me question the validity of all the other results."

Modern ML is based on induction; it is not possible to guarantee 100% accuracy with any ML agent. Training users on ML basics will help this issue. Users may need to be warned that machine learning agents 'think' differently than humans. Specifically,

similar to Dzindolet et al., users should be informed that problems that appear obvious to humans may be difficult for ML agents to detect, and vice versa (2003).

One participant reported that they felt skeptical after the ML made an error on a case similar to recent cases.

> "[I]t got a signal wrong that looked very much like the other two signals that it got right. Because the characteristics were so similar between the ones [I] agreed with and the one [I] disagreed with, it immediately makes me think there could be other signals that [I] think should be easily classified as anomalies that it's getting wrong."

The machine is able to differentiate smaller feature differences than people. As a result, the user could see the machine as a less predictable, untrustworthy agent. Basic ML training could mitigate trust decay.

Chapter 4. Extended Discussion and Conclusion

Performance Outcomes

Overall, having explanations boosted trust and understanding, with the ability to compare and contrast boosting it even more so.

Participants were able to make correct decisions roughly 60% of the time, and 70% of the time when there was an anomaly. This indicates that machines did help with anomaly detection, though we would need a no-AI condition to further investigate the impact of having AI on detection.

While we did not directly measure common ground, we still measured understanding. We did not find a significant difference in understanding between the three interfaces; however, participants ranked XAI-2 as the interface they understood best significantly more than they did the baseline. Furthermore, participants reported that having explanations increased their understanding of the interfaces when asked in free response questions. Having increased understanding could not only facilitate common ground but it could also aid in the sensemaking process.

Another important factor in the sensemaking process is informativeness –what the machine effectively conveys to the users. A highly informative interface would facilitate effective decision-making because it helps the users make better interpretations of the outputs. In this case, effective decision-making would be a participant correctly agreeing with an interface, or choosing to override an interface when its predictions are incorrect.

While the differences between interfaces were not significant, participants still agreed with the machine when it was correct most when interacting with XAI-2. Participants also had the fewest cases of disagreeing with the machine when it was right when they worked with XAI-2. Furthermore, XAI-1 had the worst performance, with participants being most likely to agree with the interface when it was incorrect or incorrectly override its prediction. This suggests that people were best able to use XAI-2's explanations to respond appropriately. This also indicates that XAI-2 had more informativeness than the other interfaces, likely due to the contrastive explanations.

Participants may have been more likely to see when XAI-2's predictions were not well aligned to the target by seeing both types of examples. In other words, participants could have seen when a classification may have been more difficult for the machine, especially if it made a mistake on a case that may have been obvious to the user.

For case types, we found that there were significant differences in performance across interfaces when participants examined cases that were not anomalies, and we found higher disagreement rates with the machine when participants examined anomalies. The different types of cases may have resulted in varying levels of caution from the participants where they may have agreed with the machine more readily when they thought the case was not an anomaly and exercised greater caution for the possibility of facing an environmental disaster.

The challenge moving forward is increasing performance. We found that XAI-2 and the baseline had no significant performance differences, so there were no detriments

to having a good explanation, but it did not necessarily improve performance. It did, however, benefit user experience.

Design Patterns

Design Pattern #1: Increasing Observability

We provided explanations where the interfaces showed visual examples from the training set that best fit a target image. Moreover, we had an interface that allowed participants to compare and contrast the images by providing examples from the other category.

People trusted and understood the interfaces that provided these explanations more than they did the baseline. Interestingly, XAI-1 had the worst performance out of all three interfaces, which indicates that simply providing these explanations are not enough. XAI-2 may have aided the sensemaking process because participants could use the other set of explanations to gain a better understanding of how the interface was making its classifications.

Design Pattern #2: Designing for Contrast

One way to prevent data overload is to show when values or images deviate or have contrasts from a baseline. Our XAI-2 interface shows examples of both healthy and anomalous pipelines. While the examples are not a set baseline, the machine was trained using images of true anomalies. Participants were able to use images from the two classes to compare differences, especially when the machine predicted an anomaly. While this may not be as beneficial when users are pressed for time in making decisions, XAI-2 had

better performance than XAI-1, and did not significantly differ from the baseline, indicating that being able to contrast images improved performance compared to XAI-1. Comparing and contrasting could help address false priming because the participants can see the contrasting examples and make more informed decisions.

Design Pattern #3: Designing to prevent fixation and the keyhole effect

Fixation occurs when users get stuck on one piece of evidence and fail to revise their hypotheses when there is new information (Woods and Hollnagel, 2006). Both of our interfaces that provided explanations output three of the best examples from the training data. Providing three examples rather than one provided the users with not only greater insight into how the ML was making predictions but it also allowed users to see alternatives rather than risk making a decision based on only one piece of evidence. In the case of XAI-2, participants were able to compare six images to determine whether a pipe was truly healthy or problematic.

Overall, we found that our explainable interfaces bolstered trust and understanding compared to the baseline. In addition, we found that there were no detriments to performance when users saw explanations. While interface designs are context dependent, supporting observability, contrast, and designing to prevent fixation can help users avoid common issues such as data overload as they make decisions and ideally help them respond more efficiently. Even so, more research would need to be conducted to further test these design patterns in other contexts and with other configurations to determine the impact of these patterns.

Beyond these patterns, our findings indicate that XAI design should include the practitioners' input, be tailored to context and to the users, include an education or training component when being introduced to a workplace, and consider compare/contrast example-based XAI but only in a less high-paced situation.

Future Design Recommendations:

We found that adding in example-based explanations can not only increase trust and understanding but that it can also help with peoples' emotional states, as we saw a shift from nervousness to understanding as a predictor of trust. Future designs should include explanations so that people perceive the machines as more observable and predictable so that they feel less nervous about using the interfaces. This should especially be the case for users in high-risk, highly technical populations, where making mistakes can result in larger consequences. For future studies, priming should be included beforehand so that users are aware that the interfaces being tested will not replace their current systems until they have been thoroughly evaluated. New designs should also include an education process. Users should have time to familiarize themselves with the new systems before using them at work. We also found that some participants had unrealistic expectations about machines having 100% accuracy, so users should also receive basic training on the capabilities and limits of the machines.

Furthermore, users should not feel overloaded by the outputs provided by the interfaces; rather, they should be able to observe the most important pieces of information and make quick decisions. Designs should also provide an overview with multiple pieces

of evidence to help prevent fixation. Users in our study were able to use several images to help them make more informed decisions. However, they were also not in a position where they needed to be able to make split-second decisions. In another context where there is a more rapid pace, example-based visuals could still be beneficial as long as the interfaces highlight changes and contrasts. Future designs could also reorganize the displays so that the most pertinent information is centered, and the less important information is in the surroundings.

Future designs should also treat classification categories differently depending on the context. In our work, we found that participants' performance differed based on whether a pipe was classified as normal or anomalous. Interactions may need to be different based on these classifications, especially if one classification will result in an emergency response or require quicker decision-making. If a machine detects an emergency, users should have settings that make it easy for them to respond to the issue. For example, users should be able to override the interface if they disagree or even include ways to notify others to maintain common ground if there is a true emergency.

Limitations

There were multiple limitations with this study. First, we only had 24 participants recruited from one company. It is unclear whether these results would extend to other technical populations. Participants interacted with the three interfaces for four trials each and the trials were randomized. It is not clear whether trust decay would continue for an extended period because only one trial had a misclassification and whether some

classifications were more difficult than others. We expect the rate of decay would be different if there were twenty-five misclassifications out of one hundred rather than one out of four. Some trials may have been easier to judge compared to others, which could have impacted whether participants agreed with the interfaces. Furthermore, participants only reported trust and understanding through self-reports and rankings; there are no other measures included that address trust with more depth. In addition, there is no dynamic interaction in this study: Trust and understanding can shift throughout interactions, but our study only measures them after participants have finished interacting with the interfaces. Lastly, to mirror their real-life workflow, participants did not receive immediate feedback as to whether each decision was correct, so it is unclear whether performance and agreement rates would have differed if there was feedback.

Future Research and Applications

Future research would entail a more generalized study with a larger sample. Specifically, participants could interact with the same interfaces for more trials to complete a simpler classification task to eliminate confusion on color scales and terminology. Participants could also have practice trials before working with the interfaces in the study to ensure that they are aware of what the interfaces can do beforehand rather than learning the capabilities partway through their interactions. In addition, the randomization would only be for the order of interacting with the interfaces to increase consistency. In our current work, we were unsure of whether some cases that were presented to participants were more difficult than others. Future research would

ensure that there are varying levels of difficulty in the trials and that each participant interacts with the same trials.

We found that having XAI shifted participants' concerns from nervousness to that of understanding. We think participants may have felt more confident in their decisions when they had explanations to help them make decisions; on the other hand, participants may have felt as though they had less control when interacting with the baseline because they had no information to evaluate. However, our current work does not examine this. Future research would investigate the impact of XAI on confidence and the need for control by comparing different levels of explanation. The study could also conduct a deeper investigation into the impact of expertise on performance with AI by including equal numbers of participants in each group with a wider range of experience levels in a domain.

Future work would include more measures for trust and performance. For trust, one possible measure beyond self-report is determining the extent to which participants agreed with each classification rather than having a binary decision. Furthermore, participants could be asked why they agreed or disagreed with the classification for each trial to gain further insight. Participants could also be asked whether they understood the machine after each trial.

For performance, one topic of interest is the impact the level of risk had on decision-making. We found that participants treated cases that were normal differently than those that were anomalous. The study could measure the time participants took to make a decision. This could then not only determine whether participants spent more

time on some cases over others, but also whether more time was spent on some interfaces over others. In emergency situations, analysts would not be able to spend too long viewing explanations, as they would need to be able to make quick decisions. If one of the interfaces that provides explanations results in longer decision-making times, it could indicate that the participants are experiencing data overload. Another condition that could be added is a no-AI condition. Participants interacted with three interfaces, but all three of them included an underlying ML algorithm. Comparing performance between the interfaces and with a no-AI for accuracy and efficiency would provide more insight into the impact of including AI into the workflow.

Conclusion

Adding an example-based explanation increased trust, understanding, and helpfulness, without reducing objective performance. Example-based explanations can shift participants from focusing on negative emotions to logical analysis while interacting with the ML.

For our highly-technical population, we found a strong positive correlation between trust and understanding. The most important feature for predicting how helpful people found the XAI's examples was understanding. The most important feature for predicting trust in all interfaces was perceived agreement.

Showing the most similar examples from multiple classes increased trust and understanding more than showing examples from only one class.

Our highly-technical population had a higher trust in the human-machine team than the ML alone. The ability for a human to override or contribute to the decision may be a more important factor than trust.

Years of experience was not significant for trust and understanding in the interface, but was significant in trust in the human-machine team and helpfulness of the XAI cases.While participants consistently ranked the XAI-2 interface as the most trusted at the end of the study, our results indicate that the ability for the human to override the ML's decision is more important to overall trust than the interface or years of experience.

The trust and understanding rankings were highly correlated, indicating that users trust the interface more when they understand it, and vice versa.

We trained Random Forest regression models to predict the human experience metrics in order to determine which features were the most important for each metric. For all the interfaces, perceived agreement was the most important feature for predicting trust in the ML, and vice versa. The most important feature for predicting the helpfulness of the explanation was how well the participant understood the explanation. For trust in the human-machine team and understanding, the most important feature for the Baseline interface differed from the XAI interfaces. When predicting trust in the human-machine team, understanding was the most important for the XAI interfaces whereas nervousness against robots/AI was the most important for the Baseline. Similarly, for predicting understanding, the XAI interfaces found helpfulness of the explanations and trust in the human-machine team to be most important, whereas the Baseline found trust in the ML's classification and perceived agreement to be most important. Overall, perceived

agreement and trust in the ML's classification are closely coupled. For XAI interfaces, trust in the human-machine team relies heavily on understanding, which is tightly coupled with helpfulness.

This study provides preliminary results that indicate that exam-ple-based explanations increase trust and understanding without having a significant negative impact on performance when used by highly-technical users. In a real setting, these users may not want to use the baseline interface regardless of accuracy because they are less likely to trust it. Thus, increasing trust is crucial for workers in technical fields to be willing to use AI.

The analyses also indicate that the quality of an explanation influences trust and performance. Specifically, interface designs should be tailored to the users and specific technical domains to ensure that explanations are beneficial; otherwise, performance may decrease. We found that trust in the human-machine team was greater than trust in the machine alone, suggesting that participants had more trust when they were involved in the decision-making process.

Furthermore, the qualitative analyses suggest that there are multiple factors that impact trust and understanding alone and that more design considerations are required in the future. Specifically, we found that for a technical population, the amount of explanation, hu-man input, interface presentation, obvious errors, and amount of human training were the key factors involved in trust. Overall, example-based explanations show promise for increasing trust and continuing to show explana-tions while training users may help mitigate further concerns with trust.

Ultimately, this research can be applied to several highly-technical domains to aid humans in their work. Although we only described four areas of research, there are many more domains in which AI can be employed. More research needs to be conducted on explainability and trust to ensure fair and safe decision-making for humans. More research also needs to be conducted on coordination to examine how XAI can improve joint activity.

Bibliography

Adadi, A., & Berrada, M. (2018). Peeking inside the black- box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., ... & Horvitz, E. (2019, May). Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems* (pp. 1-13).

Ashoori, M., & Weisz, J. D. (2019). In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes. arXiv preprint arXiv:1912.02675.

Azari, A. R., Lockhart, J. W., Liemohn, M. W., & Jia, X. (2020). Incorporating physical knowledge into machine learning for Planetary Space Physics. *Frontiers in astronomy and space sciences*, *7*, 36.

Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020, March). Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 454-464).

Cai, C. J., Jongejan, J., & Holbrook, J. (2019, March). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258-262).

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1721-1730).

Chiou, E. K., Demir, M., Buchanan, V., Corral, C. C., Endsley, M. R., Lematta, G. J., ... & McNeese, N. J. (2021). Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, 1-20.

DeepMind. (2015). *AlphaGo*. https://www.deepmind.com/research/highlighted-research/alphago

Dekker, S. (2014). *The field guide to understanding'human error'*. Ashgate Publishing, Ltd.

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

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International journal of human-computer studies, 58(6), 697-718.

Geels-Blair, K., Rice, S., & Schwark, J. (2013). Using system-wide trust theory to reveal

the contagion effects of automation false alarms and misses on compliance and reliance in a simulated aviation task. The International Journal of Aviation Psychology, 23(3), 245-266.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)* (pp. 80-89). IEEE.

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice taking from humans and machines: An fMRI and effective connectivity study. *Frontiers in Human Neuroscience*, *10*, 542.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM computing surveys (CSUR), 51(5), 1-42.

Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, *32*(3), 68-73.

Hoffman, R. R., Miller, T., & Clancey, W. J. (2022). Psychology and AI at a Crossroads: How Might Complex Systems Explain Themselves?. *The American journal of psychology*, *135*(4), 365-378.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, *55*(2), 1-38.

Kim, B., Koyejo, O., & Khanna, R. (2016, December). Examples are not enough, learn to criticize! Criticism for Interpretability. In NIPS (pp. 2280-2288).

Kim, B., Rudin, C., & Shah, J. A. (2014). The bayesian case model: A generative approach for case-based reasoning and prototype classification. In Advances in neural information processing systems (pp. 1952-1960).

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. Organizational simulation, 53, 139-184.

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a" team player" in joint human-agent activity. *IEEE Intelligent Systems*, *19*(6), 91-95.

Krishnan, R., Sivakumar, G., & Bhattacharya, P. (1999). Extracting decision trees from trained neural networks. *Pattern recognition*, *32*(12).

Krug, S. (2000). *Don't make me think!: a common sense approach to Web usability*. Pearson Education India.

Johnson, M., & Bradshaw, J. M. (2021). The role of interdependence in trust. In *Trust in human-robot interaction* (pp. 379-403). Academic Press.

Juma, C. (2016). *Innovation and its enemies: Why people resist new technologies*. Oxford University Press.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human factors, 46(1), 50-80.

Levin, T. (2021). *Tesla's full self-driving tech keeps getting fooled by the Moon,*

billboards, and Burger King signs. Business Insider. Retrieved November 28, 2022, from https://www.businessinsider.com/tesla-fsd-full-self-driving-traffic-light-fooled-moon-vid eo-2021-7?international=true&r=US&IR=T

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. Human factors, 48(2), 241-256.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*(3-4), 1-45.

O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, *64*(5), 904-938.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human factors*, *39*(2), 230-253.

Patterson, E. S., & Hoffman, R. R. (2012). Visualization framework of macrocognition functions. *Cognition, Technology & Work*, *14*, 221-227.

Rayo, M. F. (2017, September). Designing for collaborative autonomy: updating user-centered design heuristics and evaluation methods. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 61, No. 1, pp. 1091-1095). Sage CA: Los Angeles, CA: SAGE Publications.

Rayo, M. F., Fitzgerald, M. C., Gifford, R. C., Morey, D. A., Reynolds, M. E., D'Annolfo, K., & Jefferies, C. M. (2020, September). The need for machine fitness assessment: Enabling joint human-machine performance in consumer health technologies. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care* (Vol. 9, No. 1, pp. 40-42). Sage CA: Los Angeles, CA: SAGE Publications.

Rayo, M. F., Horwood, C. R., Fitzgerald, M. C., Grayson, M. R., Abdel-Rasoul, M., & Moffatt-Bruce, S. D. (2022). Situated Visual Alarm Displays Support Machine Fitness Assessment for Nonexplainable Automation. *IEEE Transactions on Human-Machine Systems*, *52*(5), 984-993.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

Rice, S. (2009). Examining single-and multiple-process theories of trust in automation. The Journal of general psychology, 136(3), 303-322.

Sarkar, S., Weyde, T., Garcez, A. D., Slabaugh, G. G., Dragicevic, S., & Percy, C. (2016, December). Accuracy and interpretability trade-offs in machine learning applied to safer gambling. In *CEUR Workshop Proceedings* (Vol. 1773). CEUR Workshop Proceedings.

Siddiqui, F., & Merrill, J. B. (2023b, June 13). 17 fatalities, 736 crashes: The shocking toll of Tesla's Autopilot. *Washington Post*. https://www.washingtonpost.com/technology/2023/06/10/tesla-autopilot-crashes-elon-musk/

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS*

Transactions on Human-Computer Interaction, *12*(3), 109-124.
Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1-39.
Woods, D. D. (2002). Escape from data overload. *Institute for Ergonomics. Ohio State University*.
Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. CRC Press.
Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020, March). How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 189-201).