Calibrating Trust in AI-Assisted Decision Making

Amy Turner, Meena Kaushik, Mu-Ti Huang, and Srikar Varanasi
UC Berkeley School of Information
{amyturner, meenakaushik, mu-ti_huang, srikar}@berkeley.edu

ABSTRACT

With the proliferation of AI products, humans and AI are increasingly working in partnership with each other to make decisions. For this type of collaboration to be successful, humans need to understand AI capability in order to effectively calibrate their trust. In these partnerships, it's critical to explain decisions and predictions in a manner that can be understood by humans in order to encourage trust calibration. The field of explainable AI is focused on integrating explainability into AI, but is geared towards making AI models more interpretable. As a result, this research often approaches explanations from a model-centric perspective, as opposed to a human-centered perspective. At the same time, industry researchers have developed guidelines to help interface designers effectively generate user-friendly explanations. However, these guidelines are typically too broad to be effective for the day-to-day work of industry designers. Our research addresses this gap through two approaches: an empirical experiment to investigate how people respond to explanations and what types of explanations are most helpful for trust calibration, and an educational resource for industry designers to help them understand what questions users might have, and how the context of use influences what explanations they may use. Findings from our experiment indicate that explanations do not always aid trust calibration, and can actually hurt it, especially in the face of novice users who have low self-competence. Our exploratory interviews and usability testing with industry designers reveal that there is a desire for a comprehensive but accessible educational resource that translates research such as our experiment and guides the design of explainable interfaces for AI products.

KEYWORDS

decision support, trust, confidence, explainable artificial intelligence, human-AI interaction, user experience

1. INTRODUCTION

Artificial Intelligence (AI) has become ubiquitous, integrating into a variety of applications. One such application is AI-assisted decision making, wherein the individual strengths of a human and AI combine in order to produce a decision outcome that is better than what either could produce alone (Zhang et al. 2020). AI can help people make decisions ranging from the mundane to the monumental; from what movie to watch or what route to take to the likelihood a defendant will reoffend or whether someone has cancer. Full automation is often undesirable, especially in high stakes scenarios such as healthcare, finance, or criminal justice, given the probabilistic nature of AI and potential input error, flaws, and biases in the underlying training data. As the performance of AI continues to improve, there is a fundamental challenge in AI-assisted decision making: balancing the powerful capabilities of AI with designing technologies that are understandable by humans (Abdul et al. 2018).

In order for these human-AI decision making partnerships to be effective, people need to know when to trust or distrust an AI's prediction, thereby forming a correct mental model of the model's error boundaries (Bansal et al. 2019). As a result, trust calibration is key to the success of AI-assisted decision making. If a person trusts AI when the AI is likely to err or distrusts AI when the AI is likely to be correct, the joint decision outcome would suffer. Poor decision outcomes can be catastrophic in high stakes scenarios. For example, ProPublica published a report that revealed alarming biases in COMPAS, a risk assessment tool used to predict the likelihood a defendant will reoffend. This is troublesome given its influence on sentencing decisions, and the higher weight it carries in the hands of judges with less experience (Angwin et al. 2016). In order to form a correct mental model of an AI model's error boundaries, a person must correctly calibrate trust on a case-by-case basis. This goal is different from building or increasing trust in AI; increasing trust does not necessarily help people distinguish cases they can trust an AI's prediction from those that they should not (Zhang et al. 2020).

There is increasing societal pressure and legislation demanding that AI systems and predictions are explained to users, such as the European Union General Data Protection Regulation (GDPR) who approved the "right to explanation" in 2016 which mandates that data subjects receive meaningful information about the logic involved in automated decision-making systems (Selbst & Powles 2017). Explanations help provide transparency, enable assessment of accountability, demonstrate fairness, and facilitate understanding (Sokol & Flach 2020). Most of the research on explainability focuses on making the models themselves interpretable, relying on a researcher's intuition on what constitutes a good explanation (Miller 2017). While creating more interpretable models is a significant contribution, there are numerous challenges to incorporating explainability in AI systems, such as the rise of "black box" models which are often unintelligible even to technical experts and the protection of intellectual property, resulting in hesitance to reveal the underlying model or algorithm. Furthermore, there is a lack of clear guidance on what type of explanations to use, when to use them, and how to create effective explainable interfaces. For example, AI transparency can have a positive impact, however if there is too much uncertainty, trust can be negatively impacted (Stowers et al., 2017). This lack of tangible guidance stems from a lack of empirical research on the human-side of explanations that examines whether users in real-world situations consider these “intelligible models” and explanation interfaces to be understandable, practical, or usable (Cheng et al. 2019).

In this paper we contribute to the human-side of explanations in two ways:

1. An empirical study of how people respond to explanations in a scenario with high uncertainty and risk;
2. An online, interactive resource that outlines a human-centered design strategy and brainstorming tool for creating explanation interfaces.

2. RELATED WORK

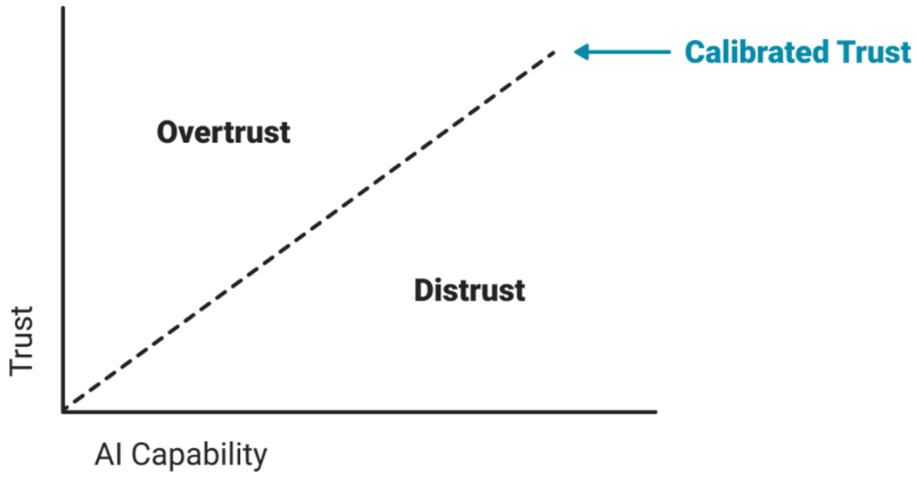
2.1 Trust

Some researchers have argued that human-AI teams function similarly to human-human teams, particularly with AI that mimics human intelligence and thereby resembles human-human trust (Kessler et al. 2017). Madhavan and Wiegmann presented a theoretical framework that synthesizes and describes the process of trust development in human-human and human-automation relationships. While people have a propensity to apply norms of human-human interpersonal interaction to human-automation interaction, there are subtle differences. In AI-assisted decision making, the human moves from being the primary decision maker to being an active teammate and sharing control with automation. As a result, decision-making processes of human-AI teams are often influenced strongly by a person’s trust in an AI team member relative to a human partner (Madhavan & Wiegmann 2007). Trust is also important in AI-assisted decision making because of its use in high stakes scenarios, wherein risk and uncertainty coexist and there is ongoing risk-taking (Cheshire 2011).

Various factors can influence use and trust in AI: trust disposition, attitude towards AI, risk, task complexity, and self-confidence (Parasuraman and Riley 1997). There are two critical elements that define the basis of trust: the focus of trust (what is to be trusted) and the type of information that describes the entity to be trusted (Lee & See 2004). This information guides expectations regarding how well the entity to be trusted can achieve the trustor’s goal. Trust in an information system, such as AI, involves expectations about the system’s predictability and reliability (Cheshire 2011). This is supported by a meta-analysis of factors that influence human-automation interactions (HAI) which revealed that the most important consideration is reliability (Kessler et al. 2017). Therefore expectations about reliability influences trust and expectations about what a system can and cannot do influences trust calibration. There are three aspects to expectation forming: external information, knowledge and understanding, and first-hand experience through a sense of control (Kocielnik et al. 2019). Expectations of system capabilities are typically set through mental models. Mental models are a person's understanding of a system and how it works.

A common mental model of AI systems is that they perform flawlessly all of the time; most people do not expect AI to behave inconsistently and imperfectly (Kocielnik et al. 2019). This can result in inattention, wherein a person fails to evaluate and monitor system performance, and overreliance. Inattention can also be caused by the high cognitive load that is often needed to monitor complex AI systems. Conversely, people expect humans to be prone to error. Therefore people will forgive a human for making a mistake but are less likely to keep using an AI if it falters, even if it outperforms humans on average (Alexander et al. 2018). Both scenarios indicate flawed partnerships, or poor trust calibration, between humans and AI.

Trust calibration refers to the correspondence between a person's trust in the AI and the AI's capabilities (Lee & Moray 1994). Overtrust, when trust exceeds the system's capabilities, leads to misuse. Misuse refers to failures that occur when people inadvertently violate critical assumptions and trust AI when they shouldn't (Lee & See 2004). Misuse can result in people relying uncritically on automation without recognizing its limitations or monitoring the automation's behavior (Parasuraman and Riley 1997). Distrust, when trust is less than the system's capabilities, leads to disuse. Disuse refers to failures that occur when people reject the capabilities of AI, failing to use it when they should (Lee & See 2004). These flawed partnerships can result in outcomes that are costly and even catastrophic.



Trust Calibration: Correspondence between Trust and AI Capability

In addition to trust calibration, specificity and resolution can mitigate misuse and disuse of AI. Resolution is a measure of how precisely a judgment of trust distinguishes between levels of automation capability. Good resolution means that the range of system capability maps onto the same range of trust. Specificity is the degree to which trust is associated with a particular component or aspect of the AI (Lee & See 2004). Specificity can be functional or temporal. Functional specificity refers to the differentiation of functions, subfunctions, and modes of AI. High functional specificity means that a person's trust reflects the specific capabilities of the AI's subfunctions and modes, while low functional specificity means that a person's trust reflects the general capabilities of the system. Temporal specificity describes the sensitivity of trust to changes in context and time that affect AI capability. High temporal specificity means that a person's trust reflects frequent fluctuations of AI capability, whereas low temporal specificity means that the trust reflects only long-term changes in AI capability. Enhancing human-AI partnerships lies in good calibration, high resolution, and high specificity of trust (Lee & See 2004).

A key question arises: how do people know whether or not to trust an AI, especially given the lack of affordances to assess trustworthiness? Trustworthiness is a characteristic or property, which can be inferred through explicit or implicit information (Cheshire 2011). A potential way to signal trustworthiness of an AI is through explainability.

2.2 Explainable AI (XAI)

Explainable AI (XAI) is a research field that aims to make AI systems more understandable (Adadi & Berrada 2018). While the term "XAI" was first coined in 2004, the problem of explainability has existed over the past decades with expert systems in the mid-1970s, Bayesian networks and artificial neural networks in the 1980s, and recommender systems in the 2000s (Abdul et al. 2018). The rise of black box models coupled with the increasing use of AI in high stakes scenarios has resulted in a recent surge of interest and focus on explainability.

Explainability is related to the concept of interpretability; interpretable systems are explainable if they can be understood by a human. There are four main needs for explainability: explain to justify a prediction, which is particularly important in cases where there is a concern for biased or discriminatory results; explain to control, enabling debugging or intervention; explain to improve; and explain to discover new information (Adadi & Berrada 2018). Various domains have a need for explanation, including transportation, healthcare, law, finance, and the military.

Explainability methods can be classified based on three criteria: (1) the complexity of interpretability, (2) the scoop of interpretability, and (3) the level of dependency from the used ML model (Adadi & Berrada 2018). Complexity related methods refers to the complexity of the model; the more complex it is, the harder it is to explain. There are two main strategies. Ante-hoc approaches use the same model for predictions and explanations. These approaches are thought to provide full transparency and are typically model-specific because they are designed for and only applicable to a specific model. Post-hoc approaches use a different model to reverse engineer the inner works of the original model and provide explanations. These approaches are thought to lighten the black box of complex models and are typically model-agnostic because they are designed to work with any type of model. Scoop related methods refer to how much of the model is being explained: the entire model behavior (global interpretability) or a specific prediction (local interpretability). Global explanations help users understand and evaluate the system. Local explanations help users examine individual cases, which can help with identifying fairness discrepancies and calibrating trust on a case-by-case basis. Finally, model related methods refer to the breadth of application of an explanation method. Model-specific interpretability refers to methods that are limited to specific model classes. Model-agnostic interpretability refers to methods that are not tied to a specific model (Adadi & Berrada 2018).

Explanations can also have various characteristics. Explanation fidelity includes soundness and completeness. Soundness refers to how truthful an explanation is with respect to the underlying predictive model. Completeness measures how well an explanation generalizes; in other words, what extent it covers the underlying predictive model (Kulesza et al. 2013). Explanations can be static or interactive. Interactive explanations accommodate a wider array of user needs and expectations. Some examples of interactivity include explanations that are reversible, collect and respond to user feedback, and allow adjustment of granularity. Explanations can have different formats. They can be delivered as a summarization (typically with statistics), visualization, text, formal argumentation, or a mixture of the above (Sokol & Flach 2020).

However, there is limited research on whether explanations and explanation interfaces are usable and practical to users. Most research measures the participant's understanding of the model which can be indicative of trust calibration; if a user understands a model, it follows that they will know when to trust it and when not to. Some research has focused on other aspects of explainability such as detecting fairness issues. From our review, few research studies have directly measured trust calibration.

Lim et al. examined how different types of explanation impacts users' experience, specifically in regard to understanding of the system and perceptions of trust and understanding of the system. The four types of explanations they tested were: Why, Why Not, What If, and How To (Lim et al. 2009). The key measures used to determine what types of intelligibility explanations would help users understand the system were task performance (in terms of task completion time) and user understanding (in terms of correctness and detail of reasons provided by the participants). The study showed that the Why and Why Not explanations, compared to the other explanations, resulted in improved understanding, increased trust in the system, and increased task performance. The authors suggested using Why explanations as the primary form of explanation and Why Not explanations as a secondary form because of the increased mental effort needed to understand the Why Not explanation (Lim et al. 2009).

Cheng et al. studied different explanation interfaces for understanding an algorithm used to make university admission decisions. The explanation types studied included (1) white-box vs. (2) black-box (showing the internal workings of the algorithm vs. not showing), and (3) interactive vs. (4) static (enabling users to explore algorithm behavior vs. not enabling). The evaluation metrics were objective understanding assessed through 3 quiz questions as well as self-reported understanding. The researchers found that both white-box explanations and interactive explanations can improve users' comprehension (Cheng et al. 2019).

Kulesza et al. tested the impact of different explanation characteristics, specifically completeness and soundness, on mental models and understanding through a qualitative study. They tested four treatments: HH (High-soundness, High-completeness), MM (Medium-soundness, Medium-completeness), HSLC (High-soundness, Low-completeness), and LSHC (Low-soundness, High-completeness). Mental models were measured through a combination of short-answer and Likert scale questions (Kulesza et al. 2013). Their findings showed that the most sound and most complete explanations (HH) were most successful at helping participants understand how the AI system worked. Participants in the HH condition also trusted their explanations more than participants in the other treatments.

Dodge et al. studied how people make judgments about fairness of machine learning systems and how different types of explanation impact that judgment. The authors utilized the COMPAS recidivism data and tested four types of explanation: two global, or system-level explanations (input influence-based and demographic-based) and two local, or instance-level explanations (sensitivity-based and case-based). The authors found that local explanations were more effective than global explanations at exposing case-specific fairness issues and that individuals' prior position on ML trust and feature fairness had significant impact on how they react to explanations (Dodge et al. 2019). Global explanations were beneficial for increasing confidence in understanding the model. The authors suggest a hybridized approach (having both global and local explanations) may be well-suited for a human-in-the-loop workflow: “using global explanations to understand and evaluate the model, and local explanations to scrutinize individual cases” (Dodge et al. 2019). They also suggested a new categorization of explanations: process oriented (how) vs. data oriented (justification). The key limitation to this work was that it was tested with crowdworkers rather than judges, who have more expertise in this field and would be the target users of this type of risk assessment tool (Dodge et al. 2019).

Zhang et al. conducted a case study of AI-assisted decision making in which humans and AI have comparable performance alone, and explores whether confidence or a local explanation can calibrate trust and improve the joint performance of the human and AI (Zhang et al. 2020). This research was conducted over two experiments, both of which used an income prediction task. The first looked at confidence score. The conditions included show vs. not show AI confidence, show vs. not show AI prediction, and full vs. partial model. They used stratified sampling of cases across confidence levels (Zhang et al. 2020). The second experiment looked at a local explanation, specifically feature importance, and only had a full-model condition. Key measures for trust included switch percentage: the percentage of trials in which the participant decided to use the AI's prediction as their final prediction and agreement percentage: the percentage of trials in which the participant's final prediction agreed with the AI's prediction. Accuracy was measured for the participant's prediction before seeing any information from the AI, the AI's prediction, and the participants final prediction after seeing the AI information. The results showed that confidence score can help calibrate trust in an AI model, whereas the local explanation had no effect on improving trust calibration and accuracy (Zhang et al. 2020).

Given the similarity to our research interests, we based our experiment design largely on this study from Zhang et al.

3. STUDY 1

For Study 1, we tested the following research question:

- What type of explanation impacts people's perceived trust of an AI model and joint accuracy of AI-assisted predictions?

For this study, we chose to focus on local explanations of an AI model. A local explanation describes the logic behind a specific, individual decision or recommendation from an AI, as opposed to the model-wide logic. We further narrowed our investigation to look at two types of local explanations: confidence score and data availability, which affect two aspects of expectation forming: external information and knowledge and understanding (Kocielnik et al. 2019). We chose to focus on confidence score because it is one of the most common types of information shown by AI systems and studied in XAI literature. This frequent use stems from confidence being a readily available metric from most AI models (known as F1 score), therefore it does not require additional engineering resources to surface to users in interfaces. While there has been prior research studies that have studied the effect of confidence on trust and trust calibration, most of this research shows confidence as a numeric value (a percentage). Given the likelihood percentages can be misinterpreted, we chose to pursue a categorical representation of confidence (high, medium, or low). We chose data availability as an explanation for the opposite reason: it is not a common explanation shown to users, however, industry guidelines (i.e. Google PAIR) and research (Lim & Dey 2009) has suggested showing inputs or data sources can help users understand AI systems. Selecting explanations from both sides of the popularity spectrum as well as from two different aspects of expectation forming allows us to understand how different types of explanations impact trust calibration.

In existing literature, there are two common ways to operationalize trust calibration: 1. overall accuracy of human participants + AI recommendations, and 2. Switch rate, or percent of the time when humans switch their responses based on the AI recommendation and explanation (Zhang et al. 2020). To answer the research question above, we tested the following hypotheses with a AI-assisted prediction task:

Hypothesis 1 (**H1**): Final accuracy

- **H1a**. Participants in the confidence explanation condition will have higher accuracy than the control group (with no explanations).
- **H1b**. Participants in the data availability explanation condition will have higher accuracy than the control group (with no explanations).

Hypothesis 2 (**H2**): Final accuracy

- Moreover, participants who experience the confidence explanations will have higher accuracy than those who experienced the data availability explanations.

Hypothesis 3 (**H3**): Switch rate

- Additionally, we hypothesize that the switch rate will be higher for the treatment groups (confidence and data availability) than for control. We do not hypothesize if one treatment group's switch rate will be higher than the other.

Hypothesis 4 (**H4**): Perceived trust

- Finally, we hypothesize that the participants who see confidence explanations will have a higher trust in the AI helper than data availability explanations, as measured by an AI Helper survey (see next section).

To test our hypotheses, we designed an AI-assisted prediction task wherein participants could achieve comparable performance to an AI model, but could perform most-optimally when calibrating appropriate trust of the AI model. We specifically selected a decision-making task in which the human participant and the AI helper would achieve best results when working together.

3.1 Experimental Design

3.1.1 Participants

For Study 1, we recruited 89 participants from Amazon Mechanical Turk. The average age of participants was 40 years old. Forty-nine percent of participants were male, and 39% were female. Education level was as follows: 48% had a 4-year degree, 20% had some college experience, 10% had a 2-year degree, 9% had a professional degree, 7% were high school graduates, and 2% had a doctorate degree. The experiment was hosted as a survey on Qualtrics, and conducted through Mechanical Turk.

3.1.2 Task and Materials

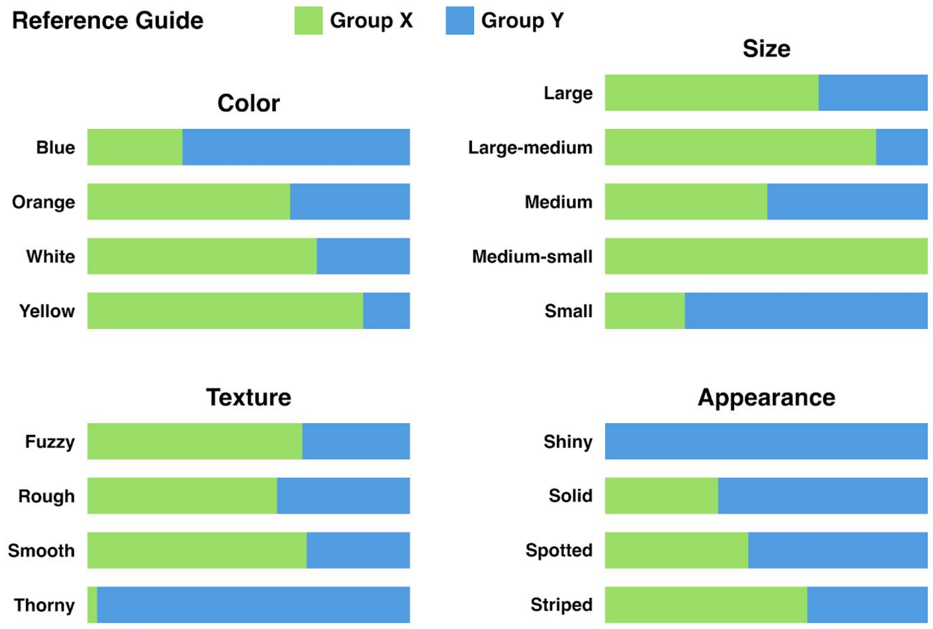
In order to test trust calibration in decision-making scenarios, we created a fictional scenario in which participants would not have previous or existing knowledge that could influence their decisions. Furthermore, this task had both uncertainty and risk. The task was highly uncertain because many plants were not obviously part of one group or the other, and the risk was personally high because payment was contingent on accuracy. To this end, we created a fictional plant, with fictional therapeutic characteristics, purported to help scientists cure cancer. These characteristics included:

- Size: large, medium-large, medium, medium-small, and small
- Texture: fuzzy, thorny, rough, and smooth
- Color: blue, white, yellow, and orange
- Appearance: spotted, stripped, solid, and shiny

The characteristics of the plants were sourced from a pre-existing data set of safe and unsafe mushrooms. The data set contains 22 characteristics for 23 species of mushrooms, and characteristics were renamed for the study. Utilizing this data set allowed us to have a ground truth to measure accuracy.

Originally, the task required participants to decide whether a specific plant was likely to be safe or poisonous. However, while pre-testing we learned that when participants were unsure whether a plant was safe or unsafe, they erred on the side of unsafe to avoid real-world negative consequences. Therefore, for the final experiment, we changed the task so that participants decided whether a plant most likely belonged to Group X or Group Y, based on its characteristics (see reference guide below), thereby mitigating risk aversion. To ensure that elevated risk was still a factor, participants were told they were receiving payment based on their accuracy, so making an informed decision had real consequences.

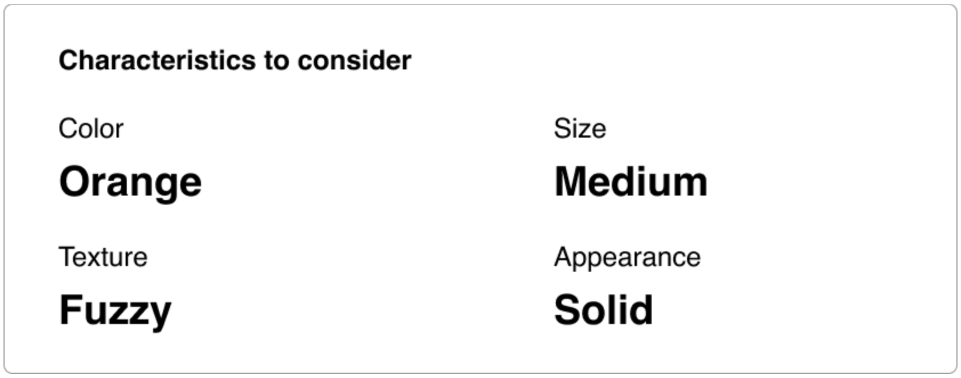
In order to aid participants with the decision-making process and avoid high cognitive load, we provided participants with the following reference guide on every question, for all treatment and control groups:



Reference guide for plant characteristics, provided to participants on each question

Treatment Tasks

Participants saw a card that showed a plant's characteristics. For each plant, participants answered "Is this plant in Group X or Group Y" twice. The first time they answered the question, referred to as their "baseline" decision going forward, they did not see the AI prediction and instead only had access to the reference guide.



Characteristics to consider

Color

Orange

Size

Medium

Texture

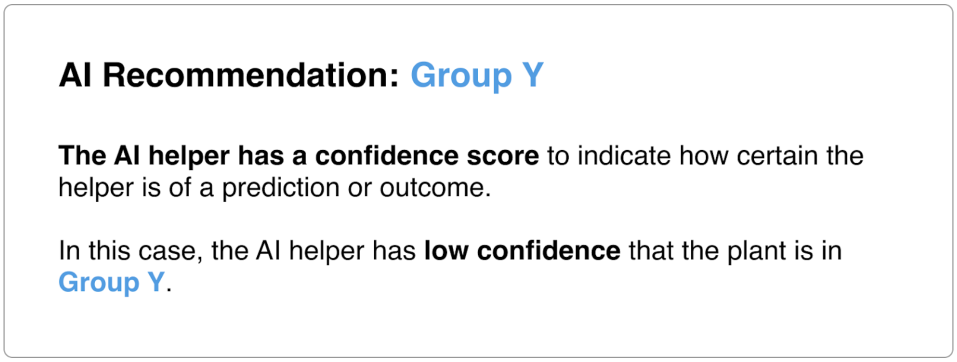
Fuzzy

Appearance

Solid

Card showing plant characteristics

The second time they answered the question, referred to as their "final decision" going forward, participants saw the AI Helper's recommendation for that specific plant, as well as additional information about the AI's recommendation (confidence or a data availability explanation). In order to measure the effect that this additional information has on accuracy, we needed participants to answer the question twice: the baseline decision **without** the AI Helper and information, and the final decision **with** the AI Helper and information. This allows us to compare any changes in accuracy due to the additional information from the AI.

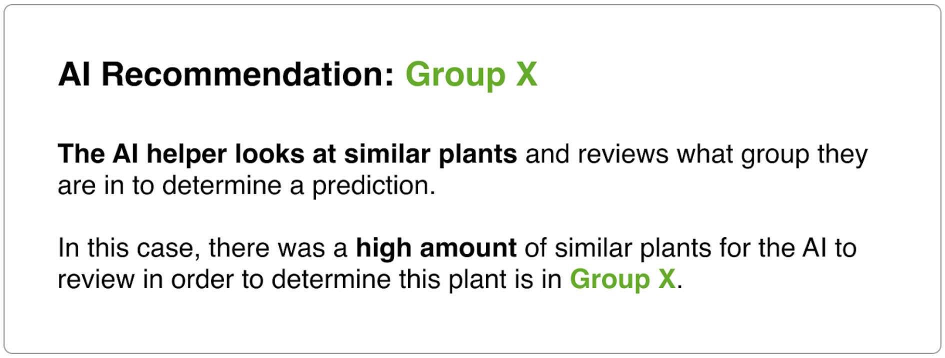


AI Recommendation: Group Y

The AI helper has a confidence score to indicate how certain the helper is of a prediction or outcome.

In this case, the AI helper has **low confidence** that the plant is in **Group Y**.

Treatment 1: AI confidence information



AI Recommendation: Group X

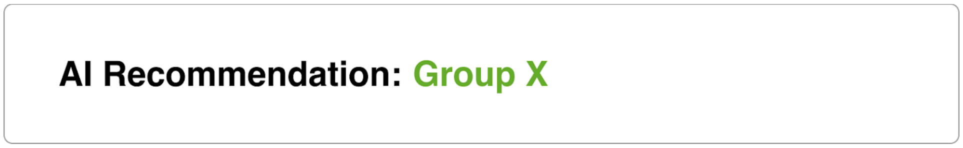
The AI helper looks at similar plants and reviews what group they are in to determine a prediction.

In this case, there was a **high amount** of similar plants for the AI to review in order to determine this plant is in **Group X**.

Treatment 2: AI data availability explanation

Control task

The control task required a baseline as well as final decision, but only presented the AI helper recommendation in the second step, without any additional information.



AI Recommendation: Group X

Control: AI recommendation without additional information

3.1.3 Design

Within-subjects

We originally conducted a within subjects experimental design, where each participant experienced a control condition, along with one of the two treatment conditions. The experiment consisted of two blocks of questions, with 30 questions total, in order; 15 control questions and 15 treatment questions. Questions in the control and treatment blocks were shown in a random order. All participants saw the same control and treatment questions.

The structure of the experiment was thus:

- Block 1: 15 control questions (AI Helper recommendation only)
- Block 2: 15 treatment questions (AI Helper recommendation **and** explanation)

Within each treatment group, there were three levels of explanations: high, medium, and low. For example, in the confidence condition the AI Helper would either have low, medium, or high level of confidence in its recommendation. For the data availability condition, the AI Helper would either have a high, medium, or low amount of similar plants to review in order to determine its recommendation. The block 2 questions were equally divided between the different levels: 5 questions that had high confidence/data availability, 5 questions that had medium confidence/data availability, and 5 questions that had low confidence/data availability.

AI model and explanations
To develop the AI Helper's recommendation and explanation, we trained a model (at 73% accuracy) on the mushroom data. The final 15 questions were not randomly selected. We chose questions that had the same level (i.e. low, medium, or high) for both confidence and data availability. This is essential to determine the effect of explanation on trust calibration, as opposed to another factor, like the characteristic of the model or the plant itself.

Between-subjects
After collecting data, we saw that there was a statistically significant difference in accuracy, within subjects, between the baseline questions in the control condition and the treatment condition. Specifically, participants were less accurate on the baseline decisions for the control questions and were more accurate on the baseline questions for the treatment questions. This could mean that the control questions were harder than the treatment questions. Therefore, any changes we observe in the treatment conditions may be due to the difference in question difficulty, instead of our manipulation. To account for this, we conducted a between-subjects control group. The control condition was identical to the treatment groups, except that participants did not see any AI Helper explanation.

Thus, the structure of the experiment was:

- Block 1: 15 control questions (AI Helper recommendation only)
- Block 2: 15 control questions (AI Helper recommendation only)

3.1.4 Dependent Measures

Final Accuracy

Trust calibration, in part, is operationalized as the accuracy of a joint human-AI decision. In both study 1 and study 2, this means average accuracy achieved by participants when making final decisions with information from an AI. If participants know when to trust the AI recommendation and when not to, it follows that they are likely to be more accurate overall.

Switch Rate

Trust calibration is also commonly operationalized as the percent rate at which humans switch their decisions to be in line with the AI decision. Further, we examine implications of this switching on overall accuracy. For example, a higher rate of switching might lead to better accuracy in cases where the AI Helper's recommendation is correct. However, it could also lead to lower accuracy if participants are blindly trusting the AI Helper in cases where it's recommendation is incorrect. Trust calibration is occuring in the former example, but not the latter. In both study 1 and study 2, this means targeting questions on which participants' initial response was different from the AI Helper recommendation, and the participant switched their answer on their final decision.

Disagreement accuracy

Disagreement accuracy is the accuracy percentage of the questions where the AI helper and participants disagreed but the participants were right in their final decision. This would mean they calibrated correctly when to trust the AI helper and when to not trust it.

Trust in the AI Helper

Specific trust of the AI Helper used in the study was assessed using the same self-report scales adapted from the Merrit (2011) Trust Scale. The adaptations included substituting their reference to "AWD" to "AI Helper", with additional text in the instructions making it clear that participants should respond to the questions keeping in mind the AI Helper they interacted with during the study.

3.1.5 Additional Measures

Explanation Satisfaction

Satisfaction with the AI explanations was assessed using a survey was created by Hoffman et al.. (2018). They argue that measuring what explanations users choose (through attitudes like satisfaction) are critical to consider when building trust in AI systems; for example it's possible that users may choose a shorter, less exhaustive explanation as preferable. Participants responded on a 5-point Likert-style scale from 1 (strongly disagree) to 5 (strongly agree). An example item is “The explanations from the AI helper are useful to my goals.”

Trust in AI Systems
General trust of AI systems was assessed using self-report scales adapted from the Merrit (2011) Trust Scale. The adaptations included substituting their reference to “AWD” to “AI system” with additional text in the instructions giving examples of what AI systems could be (smart assistants like Google Home, Siri, as well as recommendations like Netflix and GPS navigation). Participants responded on a 7-point Likert-style scale from 1 (strongly disagree) to 7 (strongly agree). An example item is “I believe AI systems are competent performers.”

3.2 Study 1 Results

We used independent samples t-test to check if there is any significant difference between the treatment groups and control.

3.2.1 Final Accuracy

Final accuracy refers to the percentage of questions on which the participant was correct on their final decision.

Condition	**Mean**
Control	77.85%
AI Confidence	76.09%
Data Availability	72.00%

Condition	**P-value**
AI Confidence vs Control	0.38
Data Availability vs Control	0.06
Data Availability vs AI Confidence	0.06

We observed that there was no significant difference between AI confidence and control groups when looking at overall accuracy. We can reject the H1a hypothesis.

We also observed a borderline significance between the data availability condition and AI confidence condition for final accuracy, meaning participants were less accurate in the data availability condition when compared to the AI confidence condition. We also found a borderline significance between data availability explanations and control explanations, meaning participants in the data availability condition were less accurate than in those in the control condition. These findings mean that the H2 hypothesis can be partially supported.

3.2.2 Switch rate

Switch rate refers to the percentage of questions where participants were originally in disagreement with the AI helper and changed their answer on the final decision.

Condition	**Mean**
Control	10.4%
AI Confidence	13.3%
Data Availability	15.5%

Condition	**P-value**
AI Confidence vs Control	0.42
Data Availability vs Control	0.12
Data Availability vs AI Confidence	0.56

We observed that there was no significant switch rate between AI confidence and control. We also found no significant difference between data availability and AI confidence. However, we found borderline significance between data availability and control, meaning that participants in the data availability condition switched more often than those in the control condition. H3 was partially supported.

3.2.2 Trust in the AI Helper

Trust in the AI helper was measured using a scale from 1 to 7.

Condition	**Mean**
Control	5.67
AI Confidence	5.24
Data Availability	5.41

Condition	**P-value**
AI Confidence vs Control	0.11
Data Availability vs Control	0.30
Data Availability vs AI Confidence	0.56

We observed that there was no significant difference in AI trust between data availability and control. We also found no significant difference between data availability and AI confidence. We found borderline significance between AI confidence and control, meaning participants trusted the AI confidence explanation less than participants in the control condition. H4 can be rejected.

3.2.3 Disagreement accuracy

Disagreement accuracy is the percentage of the questions where the AI helper and participants disagreed but the participants were right in their final decision.

Condition	**Mean**
Control	66.6%
AI Confidence	65.3%
Data Availability	52.3%

Condition	**P-value**
AI Confidence vs Control	0.85
Data Availability vs Control	0.02
Data Availability vs AI Confidence	0.04

We observed that there was no significant difference in disagreement accuracy between AI confidence and control conditions. However, we found significant differences between both data availability and AI confidence versus control conditions. This means that participants in both the control and AI confidence conditions were significantly more likely to have correct answers in the case of disagreement compared to the Data Availability condition.

4. STUDY 2

Given the results from Study 1, we hypothesized that participants did not find the need to rely on the AI Helper because their accuracy alone (70-72%) was essentially the same as the model's accuracy (73%). Therefore, would explanations become more salient if participants did not feel as competent in the decision-making task? Existing research shows that when users feel less competent they are more likely to rely on AI (Parasuraman & Riley 1997).

Therefore, our research question for study 2 is:

- When humans have low self-competence in their ability to complete a task, how do explanations of AI predictions impact people's perceived trust of an AI model and joint accuracy of AI-assisted predictions?

To answer the research question above, we conducted another experiment that was identical to Study 1, but we manipulated participant self-competence to be low.

*Hypothesis 1 (**H1**): Final accuracy*

- **H1a**. Participants in the confidence explanation condition will have higher accuracy than the control group (with no explanations).
- **H1b**. Participants in the data availability explanation condition will have higher accuracy than the control group (with no explanations).

*Hypothesis 2 (**H2**): Final accuracy*

- Moreover, participants who experience the confidence explanations will have higher accuracy than those who experienced the data availability explanations.

*Hypothesis 3 (**H3**): Switch rate*

- Additionally, we hypothesize that the switch rate will be higher for the treatment groups (confidence and data availability) than for control. We do not hypothesize if one treatment group's switch rate will be higher than the other.

*Hypothesis 4 (**H4**): Perceived trust*

- Finally, we hypothesize that the participants who see confidence explanations will have a higher trust in the AI helper than data availability explanations, as measured by an AI Helper survey (see next section).

To test our hypotheses, we utilized the same AI-assisted prediction task from Study 1 wherein participants could achieve comparable performance to an AI model, but could perform most-optimally when calibrating appropriate trust of the AI model. We hypothesized that the switch rate in study 2 will be higher than the switch rate in study 1 for both treatment groups given the competence manipulation.

4.1 Experimental Design

4.1.1 Participants

For Study 2, we recruited 92 participants from Amazon Mechanical Turk. The average age of participants was 41 years old. Forty-four percent of participants were male, and 47% were female, with one participant declining to answer. Education level was as follows: 42% had a 4-year degree, 18% had some college experience, 15% had a 2-year degree, 5% had a professional degree, 18% were high school graduates, and 0% had a doctorate degree. The experiment was hosted as a survey on Qualtrics, and conducted through Amazon Mechanical Turk.

4.1.2 Task and materials

The scenario of the task in study 2 was identical to study 1 (i.e. fictional plant, participants needed to decide if the plant was part of Group X or Group Y, based on information from a reference guide containing plant characteristics).

The treatment tasks were identical to the treatment tasks in Study 1.

4.1.3 Design

Between-subjects

For study 2, we conducted between-subjects experimental design, meaning there were 2 treatment groups and a control group. The experiment consisted of two blocks of questions, with 30 questions total, in order; 15 baseline questions and 15 treatment questions. Questions in the control and treatment blocks were shown in a random order. All participants saw the same control and treatment questions.

There were two notable differences between Study 1 and Study 2. First, participants experienced a manipulation that lowered their self-confidence in the task, halfway through. This manipulation stated that their accuracy on the previous 15 questions was "below average", and they were shown a bell curve, with their accuracy presented in the lower half. Second, for the first block of 15 questions, participants did not see the AI Helper. By manipulating participant self-confidence **before** introducing the AI Helper, we removed the possibility of undermining participant trust in the AI Helper.

Treatment design

The treatment conditions were identical to Study 1. Taken with the information above, the structure of the treatment experiments was:

- Block 1: 15 baseline questions (no AI Helper nor explanation)
- Block 2: 15 treatment questions (AI Helper recommendation **and** explanation)

Again, within the treatment questions, there were three levels of explanations (low, medium, high), with 5 questions of each type.

Control design

The only difference in the control condition was that participants only saw the AI Helper recommendation in the second block of questions; they did not see an explanation.

The structure of the control condition was:

- Block 1: 15 baseline questions (no AI Helper nor explanation)
- Block 2: 15 control questions (AI Helper recommendation only)

4.1.4 Measures

The measures for Study 2 are the same as Study 1: final accuracy, switch rate, disagreement accuracy, trust in AI helper, explanation satisfaction, and trust in AI systems.

4.2 Study 2 Results

We used independent samples t-test to check if there is any significant difference between the treatment groups and control.

4.2.1 Final Accuracy

Final accuracy refers to the percentage of questions on which the participant was correct.

Condition	**Mean**
Control	75.0%
AI Confidence	75.1%
Data Availability	73.0%

Condition	**P-value**
AI Confidence vs Control	0.94
Data Availability vs Control	0.58
Data Availability vs AI Confidence	0.56

We observed that there was no significant difference between AI confidence and control groups. We can reject the H1a hypothesis.

We also observed that there was no significant difference between Data Availability and control as well as AI confidence groups. We can reject the H1b and H2 hypotheses.

4.2.2 Switch rate

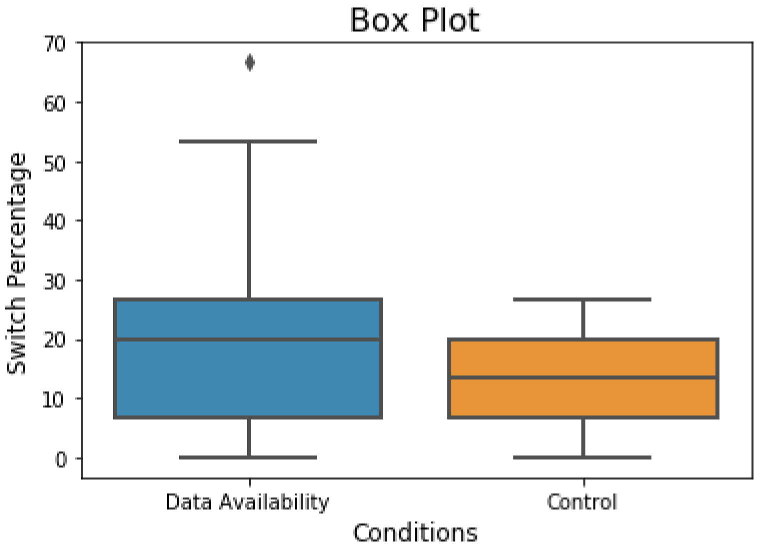
Switch rate refers to the percentage of questions where participants were originally in disagreement with the AI helper and changed their answer on the final decision.

Condition	**Mean**
Control	11.0%
AI Confidence	12.3%
Data Availability	19.0%

Condition	**P-value**
AI Confidence vs Control	0.42
Data Availability vs Control	0.04
Data Availability vs AI Confidence	0.02

We observed that there was no significant switch rate between AI confidence and control. We also found no significant difference between data availability and AI confidence.

We found a significant difference in switch rate between data availability and control, meaning participants in the data availability condition who originally disagreed with the AI helper recommendation, more often switched their answer on the final decision, when compared to the control condition. H3 was fully supported. We also found a significant difference in switch rate between data availability and AI confidence. Participants switched at a higher rate in the data Availability condition, when compared to the AI confidence condition..



Box plot for switch percentage between the conditions

4.2.3 Trust in the AI Helper

Trust in the AI Helper was measured on a scale from 1 to 7.

Condition	**Mean**
Control	4.98
AI Confidence	4.96
Data Availability	5.3

Condition	**P-value**
AI Confidence vs Control	0.93
Data Availability vs Control	0.26
Data Availability vs AI Confidence	0.28

We observed that there was no significant difference in AI trust between Data availability and control. We also found no significant difference between Data availability and AI confidence as well as control. H4 can be rejected.

4.2.4 Disagreement accuracy

Disagreement accuracy is the percentage of questions where the AI helper and participants disagreed but the participants were right in their final decision.

Condition	**Mean**
Control	54.9%
AI Confidence	65.3%
Data Availability	56.3%

Condition	**P-value**
AI Confidence vs Control	0.33
Data Availability vs Control	0.34
Data Availability vs AI Confidence	0.85

We observed that there was no significant difference in disagreement accuracy between AI confidence and control conditions. We also observed no significant difference between data Availability and AI confidence/control group.

5. EXPERIMENT DISCUSSION

Although our results did not support many of our hypotheses, they did provide insight into how people approach decision-making with artificial intelligence. We conclude that different types of explanations do impact how people behave towards AI decision-making partnerships. Certain explanations are more effective than others, while others can be detrimental to trust calibration. How these different types of explanations are acted on can also depend on how competent users feel when interacting with AI.

Confidence score led to higher accuracy than the data availability explanation, but confidence score had the same accuracy as the control group. These results may indicate that explanations aren't as critical in certain situations. The scenario tested in the experiment was one in which the AI model and participants performed at about the same accuracy rate. Furthermore, participants were told they have the same information as the AI. As a result, it could be that in scenarios where the AI seems to have the same competence and doesn't offer any additional information to the human, people do not pay attention to explanations from the AI. Another potential explanation for the increase in accuracy compared to data availability is that confidence score is a more direct representation of reliability, whereas the data availability explanation is more indirect, requiring the participant to interpret the explanation and infer the reliability of the prediction. We suspect that data availability may have led to increased understanding of the system as it provided more insight into the rationale behind the prediction compared to the confidence value, however further research should confirm this hypothesis.

Interestingly, the data availability explanation seemed to inhibit trust calibration. After participants saw the data availability explanation, in particular, they were significantly more likely to switch their final decision to match the AI recommendation, when compared to confidence score and control. The tendency to switch was magnified when participants saw the low self-competence manipulation. We saw that in Study 1 the switch rate was borderline significant, however when lowering participants' self-competence in Study 2, the switch rate became significant. Lowering self-competence appears to stimulate people's willingness and inclination to lean on the AI recommendations, at least when seeing certain types of explanations. We hypothesize that data availability may have created perceptions that the system is more capable than it is, resulting in overtrust or misuse. This desire to rely on AI recommendations occurs despite causing harm to the participant's accuracy (as we saw in Study 1). This finding has powerful implications when considering the multitude of instances where AI is providing recommendations in high stakes situations. If users don't feel competent, they might overly rely on AI to the detriment of themselves and others affected by their decisions.

The implications from the switch rate finding is also relevant to industry product development. We hypothesized that switch rate is indicative of indecision, and a perceived sense of low self-competence. Users with low self-competence are often novice users who may misuse AI, relying on it uncritically and trusting it in cases they shouldn't. This finding is especially relevant in high stakes scenarios, where AI-assisted decision outcomes have a significant impact on human lives. One example for this is a risk-recidivism tool developed by a private company that uses AI to assign risk scores to defendants in order to predict how likely they are to reoffend. Angwin et al. (2016), from Propublica, revealed that the machine learning underlying the system was biased, wrongly labelling black defendants as future criminals at a rate twice that of white defendants. One of the judges interviewed in this report stated that the scores were more helpful to them when they were a new judge. This leads to the plausible scenario that an inexperienced judge who is a novice user of this tool, may overly rely on this biased system with ineffective explanations when making crucial decisions about people's futures. In this recidivism example, often judges do not understand the bias inherent in the AI model, but given an appropriate explanation that makes this bias clear, they could better detect issues of fairness and make more informed decisions, calibrating their trust (Dodge et al. 2019). However, if novice judges are shown explanations that are not helpful for trust calibration and may inflate the perception of capability of the system, their misuse of the system can increase. This implication demonstrates why understanding a user's context, questions, and domain is crucial for designing effective explainable interfaces.

6. LIMITATIONS AND FUTURE WORK

There were certain limitations for our work. One limitation was in the design of the scenario and model accuracy. As mentioned previously, participants had a similar accuracy rate of the trained AI model. This potentially caused them to not rely on the explanations in Study 1, therefore limiting the necessity of trusting the AI recommendation. Future work could examine what explanations are helpful for trust calibration when model accuracy is higher. Another limitation of our study is that participants were not experts in the domain. In a true human-AI partnership, the human has unique, complementary knowledge to contribute. This may explain the lack of results we saw in final accuracy, which can depend on the ability of the human to bring in unique knowledge to complement the AI's errors. In order to remove the risk of choosing a scenario participants had previous experience with, we opted for a fictional scenario. However, future work could investigate the effect of explanations in real-world situations where humans are domain experts, versus novices.

Additionally, we used text-based explanations with categorical levels. It is possible other types of design such as visualizations or interactivity for the explanation may show different results. As discussed in the introduction, there are many ways of approaching a single explanation type. For example, many products show confidence as a percentage, or a letter grade. Therefore, we were limited in testing one expression of each information type in our experiment.

There were many directions our research could take which future work can address. Starting with the experimental design, there is potential to follow-up on how the sequence of events could change people's trust calibration. For example, if participants saw the AI recommendation first, before making an initial decision, the explanations could have a different effect on trust calibration than our study observed. Another interesting avenue of research could be to look at the effect of risk on decision-making with AI. Research shows that under conditions with high risk, trust becomes essential, and in some cases, can lead to an increased reliance on AI (Lyons & Stokes 2012). Therefore, raising the level of risk along with showing explanations could affect trust calibration.

Future work could further explore our specific findings in this paper from a qualitative perspective. For instance, a qualitative study could delve deeper into why data availability explanations in particular led participants to lower accuracy and higher switch rates. There is also potential for future research to investigate the effect of other explanations, in addition to the ones in this study. Specifically, counterfactual and example-based explanations are gaining interest in the Explainable AI field (Liao et al. 2020). However, there is limited research on the effect of these types of explanations on trust calibration.

7. EXPLAINABLE AI RESOURCE

7.1 Background and Related Work

With increasing societal pressure for explanations and legislation like GDPR's right to explanations, it's no longer a question of 'if we should explain AI models', but how and when we explain. However, the context of use is essential to take into account when designing explanations. As we found from our experiment, there are contexts when perhaps an elaborate explanation works against trust calibration, but there are certainly contexts in which explanations are essential for calibrating trust. With this nuance and complexity in mind, we created a resource for industry designers with the goal of helping them create effective explanations for AI systems.

The role of creating effective explanations often falls to UX design practitioners who bridge the gap between the people using the product (“users”) and the product team (including data science and engineers). Their challenge is to create design solutions accounting for demands and constraints coming from users, the product team, and other stakeholders. Both industry and academic researchers have conducted empirical research and developed insights for explainable AI, but each have their limitations.

Among the leaders in the industry space are IBM Research, Google’s People + AI, and Microsoft. IBM produced *Design for AI,* which is a broad set of design guidelines. Although unstated, its intended audience appears primarily to be design practitioners for the purpose of educating them broadly on what artificial intelligence is, and the implications this holds for design. This guide has an extremely broad approach to the field; starting with a broad introduction, how to use design thinking within a team working on an AI system, followed by fundamentals of artificial intelligence. While the broad approach is initially helpful, it leads designers wanting more concrete approaches for designing interfaces involving AI. The guide does get specific in only one section focused on designing chatbots. For example, it provides guidance on practical applications for bots (when to use them, when not to use them), as well as components to consider when building the bot itself (conversational understanding, intent, repair). This level of granularity is what is lacking in many industry best practices, so *Design For AI* would benefit from expanding to include this level of granularity for other AI products.

The People + AI research team at Google takes a different approach. They produced a guidebook that offers very broad guidance for developing AI in ‘human-centered’ ways. It is divided into six chapters, each focusing on a specific consideration (i.e. explainability + trust, mental models, etc.). The guidebook is principle-led, and seems intentionally broad so that product teams can take the concepts and apply them to specific use-cases. However, the guidebook seeks to make these principles more tangible for the end user by providing examples from Google’s product suite, as well as short worksheets after every section. While the guidebook attempts to deliver a succinct and digestible ‘how to’ for designers, it lacks the granularity necessary to be a useful resource for designers.

Microsoft Research proposed and tested 18 design guidelines for AI systems, in an effort to offer practitioners with generally applicable heuristics by which to evaluate observable “AI infused” interfaces (Amershi et al. 2019). The 18 guidelines are divided between stages of user interaction in which each is most relevant (i.e. initial interaction, during a single interaction, when something goes wrong, and across interactions over time), with accompanying examples from products (i.e. voice assistants, e-commerce, autocomplete, recommendations, etc). The guidelines are coupled with examples of how they may be applied to various products. While comprehensive, only a few guidelines refer to explainability: (1) Make clear what the system can do, (2) Make clear how well the system can do what it can do, (3) Show contextually relevant information, and (4) Make clear why the system did what it did (Amershi et al. 2019). These guidelines provide a strong starting point but lack actionable guidance on how to apply it to different products and scenarios than those listed.

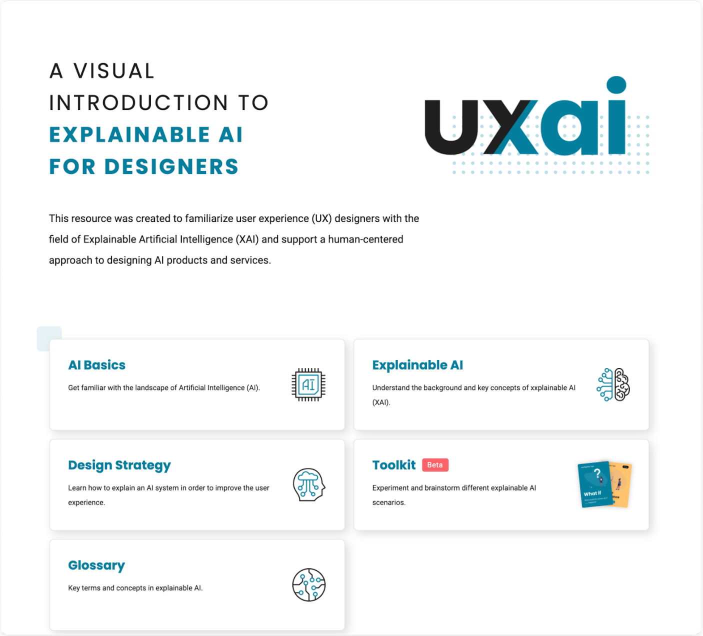
Academic research is often at the granularity that designers desire, however the focus areas and academic paper format likely impede non-academic audiences, and more specifically UX practitioners, from consuming the content. A lot of the work in AI and ML communities tend to suffer from a lack of usability, practical interpretability, and efficacy on real users (Abdul et al. 2018). The model-centric approach is evident in Explainable AI literature, resulting in papers laden with AI and ML jargon that reduces the comprehensibility for those who are not data scientists, despite having insights applicable to UX practitioners. Liao et al. explored explainability from a user's perspective by creating a question bank and interviewing current UX professionals on their experience encountering these questions from users. Their findings provided concrete strategies for exploring various types of questions and explanations that can be used to aid explainability in AI products, however, some of these strategies can be difficult to envision (Liao et al. 2020). From our interviews with professionals, we found that searching academic papers at the start of a project is not typical for industry designers, locking away essential, helpful knowledge and highlighting a need for a more approachable format.

7.2 Design Approach

Little is known about how to put techniques from research into practice, especially when bridging user needs and technical capabilities (Liao et al. 2020). The need for a resource that can not only connect academic research to industry but also provide more actionable insights was voiced in our early interviews with designers who work on AI products or guidelines. Therefore our primary goal for the website was to create a resource that informs and empowers UX designers when designing explainable interfaces. In deciding what content to feature on the website, we synthesized findings from a wide array of academic and industry resources and emphasized common threads. We leveraged the granularity of academic research and the practicality of industry resources to create an accessible and usable resource for UX designers and AI product teams to experiment with, break, and collaborate on creating explainable interfaces.

7.3 Website

Our online resource seeks to surface critically informative, granular information otherwise buried in academic papers, in an approachable way that is more inline with current industry guidelines. The website is divided into sections, working from a broad overview of AI to a tangible brainstorming tool.



A VISUAL INTRODUCTION TO **EXPLAINABLE AI FOR DESIGNERS**

This resource was created to familiarize user experience (UX) designers with the field of Explainable Artificial Intelligence (XAI) and support a human-centered approach to designing AI products and services.

AI Basics

Get familiar with the landscape of Artificial Intelligence (AI).

Explainable AI

Understand the background and key concepts of xxplainable AI (XAI).

Design Strategy

Learn how to explain an AI system in order to improve the user experience.

Toolkit Beta

Experiment and brainstorm different explainable AI scenarios.

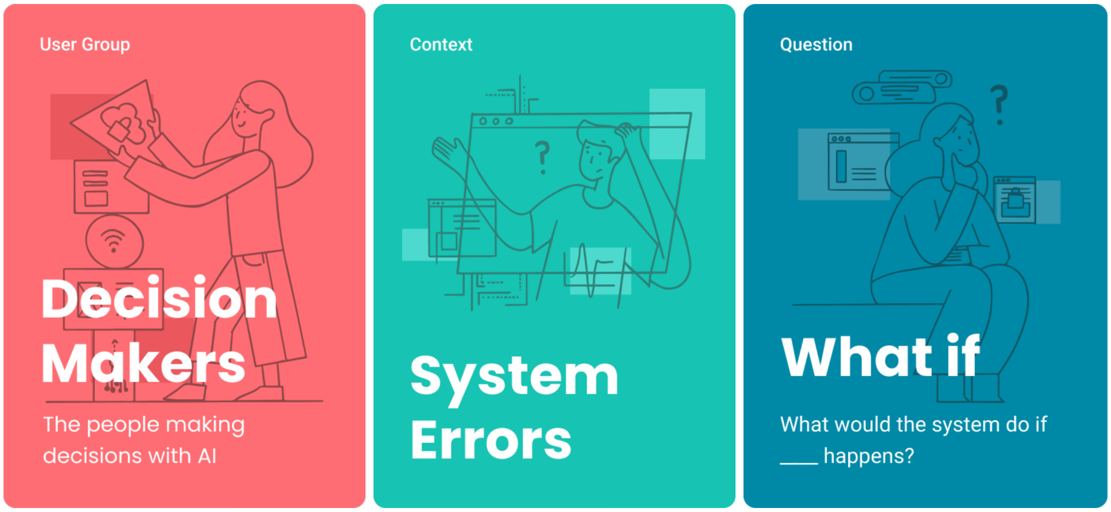
Glossary

Key terms and concepts in explainable AI.

Homepage of uxai.design

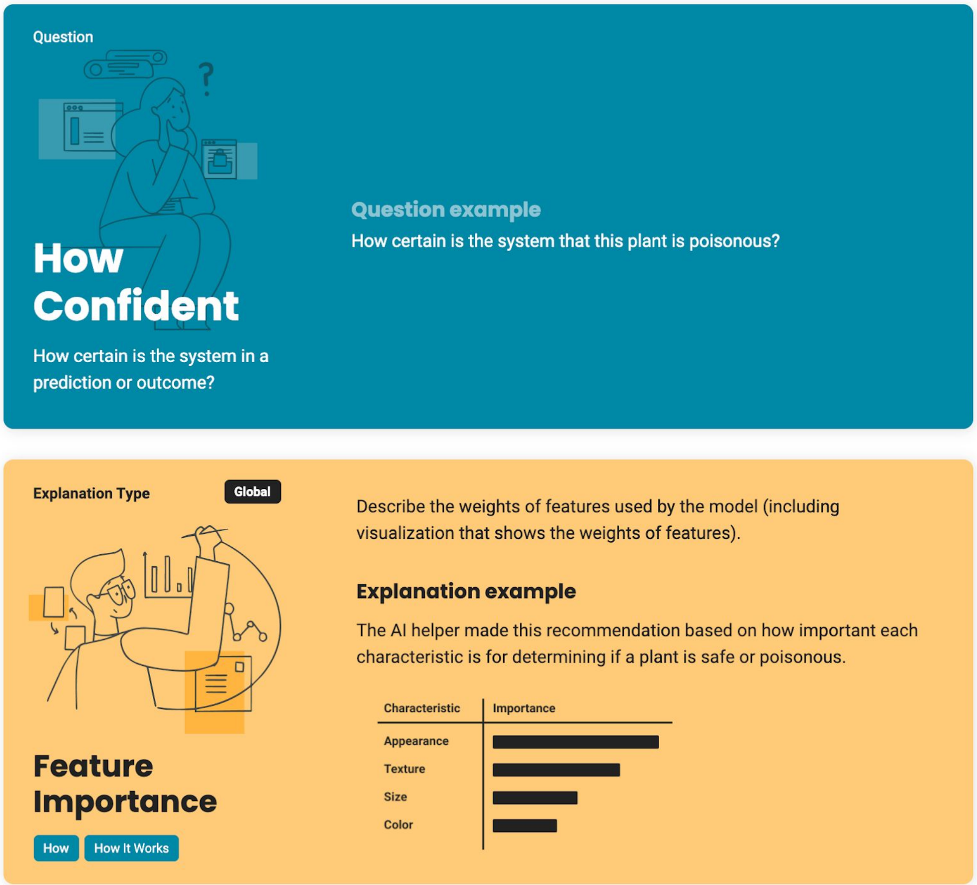
The AI Basics section covers how AI functions on logic to different types of machine learning and an overview of AI applications and decision aids. Given most products currently use, or likely will use AI, it is increasingly important for designers to be familiar with basic AI concepts.

Next, we introduce Explainable AI with a case study of why explanations matter, accompanied by industry resources. The design strategy section that follows, outlines our proposed approach, based on an extensive literature review demonstrated in this paper. This section starts by recommending designers identify and understand their users, the context in which they may need explanations, and what types of questions they are likely to have about the AI system. Then it moves on to technical considerations in asking and answering questions and ends with methods to assess and evaluate these explanations.


User Group
Decision Makers
The people making decisions with AI
Context
System Errors
Question
What if
What would the system do if ____ happens?

Example cards from the design strategy section, giving insight into different user groups, contexts, questions, and explanation types (not pictured) that designers should consider in the process.

Finally, we unveil the brainstorming cards, where we introduce the AI decision aid scenario that underpins all of the example explanations. The decision aid scenario is similar to the experiment in this paper, where an AI helper recommends a plant is safe or poisonous, but the user makes the final decision. While only one scenario is provided on the website, this tool can help designers brainstorm how to create effective explainable interfaces for various scenarios with their cross-functional AI teams. We finish the website with a glossary of terms and concepts in AI and explainability.



Question

How Confident

How certain is the system in a prediction or outcome?

Question example

How certain is the system that this plant is poisonous?

Explanation Type

Global

Describe the weights of features used by the model (including visualization that shows the weights of features).

Explanation example

The AI helper made this recommendation based on how important each characteristic is for determining if a plant is safe or poisonous.

Feature Importance

How

How It Works

Example expanded brainstorming card, pairing a question with a potential explanation type (feature importance).

7.4 Usability testing

The purpose of usability testing was to collect feedback on the usability interactive elements, the legibility of the content (appropriate and understandable terms, etc.), the logic of the information architecture, and whether or not participants would utilize the website as a tool in their own practice. Specifically, we wanted to understand if:

1. Designers could learn about AI and XAI through our resource
2. Our toolkit would be a useful addition to their workplace

We employed qualitative usability testing as a means to address the rationale above. The focus of qualitative usability testing is to understand how participants perceive the user interface, and then deliberately probe to understand why. We asked each participant several simple prompts throughout each section of the website, then allowed participants to explore the prototype naturally, while thinking out loud, to discover how intuitive the features and language were.

Participants included 3 UX designers with experience designing AI products at large technology companies. One participant had additional experience of creating AI design guidelines, specifically for explainability and trust. They were selected from the researchers personal networks.

From testing, we gathered many insights that were incorporated into the design of the current website, three of which are explored below. First, we changed language throughout to be less technical, and in areas where technicality was needed, we clearly defined those terms. This change was based on the finding that our language was not accessible enough. For example, one participant said "I don't know what Explainable AI -- is that a noun?", which clearly indicated they didn't understand the terminology that is central to our resource. Another insight was that participants were looking for the combination of a hook, explaining why Explainable AI matters, as well as a clear understanding of how they will personally benefit from our website. Therefore, in the updated version of the website, we include a case study of the self-driving car accident in Tempe, Arizona, as well as a brief history to illustrate the impact XAI can have. The final impactful insight was that although all participants were excited when they discovered the resource page, they all requested that this be front and center since these tools are what matter most for designers. As one participant said, "you are burying the lead" by having the cards come last. To address this, we included the brainstorming cards earlier in the website, integrated throughout the design strategy section.

7.5 Future Work

Future work includes more usability testing with a broader audience. In our initial study we chose designers who work on AI products at large technology companies, however, we are curious to test our site with designers from smaller start-ups as well as designers with little to no experience with AI. We are also interested in hosting an ideation workshop with cross-functional teams to study the effectiveness of our brainstorming tool: what works, what doesn't work, and what is missing. Finally, we intend to refine the website to include more visual aids and interactivity and add scenario cards to the toolkit to facilitate learning.

CONCLUSION

Our project contributes to the ongoing conversation about accountability, transparency, and explainability of artificial intelligence in several ways. First, we provided empirical evidence whether explanation interfaces are actually understandable, usable, or practical in real-world situations. Specifically, we uncovered an instance wherein an explanation can nudge people in the wrong direction, which degrades trust calibration and demonstrates how important contextual understanding is for designers of AI products. It challenges the notion that any explanation is better than no explanation for trust calibration. Second, we built a bridge between research and industry, turning research such as our experiment into a design exercise and inviting UX professionals to engage in this dialogue with their cross-functional teams. We also provided tools for designers to explore how to integrate explainable AI in their work. Our contributions to the human-side of explanation is important especially given that in AI-assisted decision making, we are not designing products -- we are designing relationships.

ACKNOWLEDGEMENTS

We thank all our anonymous participants. We also thank our advisor, Professor Coye Cheshire, for his guidance, support, and patience.

REFERENCES

1. Abdul, A., Vermeulen, J., Wang, D., Lim, B.Y., & Kankanhalli, M.S. (2018). Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. *CHI '18.*

2. Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6,* 52138-52160.

3. Alexander, V., Blinder, C., & Zak, P.J. (2018). Why trust an algorithm? Performance, cognition, and neurophysiology. *Comput. Hum. Behav., 89,* 279-288.

4. Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., … Bennett, P. N. (2019). Guidelines for Human-AI Interactio*n. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI 19.* doi: 10.1145/3290605.3300233

5. Angwin, J., Larson, J., Mattu, S., and Kirthner, L. Machine Bias. *Propublica, https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

6. Bansal, G., Nushi, B., Kamar, E., Weld, D., Lasecki, W., & Horvitz, E. (2019). Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *AAAI Conference on Artificial Intelligence.*

7. Bucher, T. (2017). The algorithmic imaginary: exploring the ordinary affects of Facebook algorithms. *Information, Communication & Society, 20,* 30 - 44.

8. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. CHI '19.

9. Cheshire, C. (2011) Online Trust, Trustworthiness, or Assurance?. *In Daedalus 140* (4): 49-58.

10. Cutler, Adam, Milena Pribic, Lawrence Himphrey, (n.d.) *Everyday Ethics for Artificial Intelligence*. IBM, Retrieved from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

11. Dodge J., Liao Q. V., Zhang Y., Bellamy R. K. E., Dugan C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19.* doi: 10.1177/1541931213601808

12. Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. *arXiv: Machine Learning.*

13. Google. (n.d.). People AI Guidebook. Retrieved from https://pair.withgoogle.com/chapter/explainability-trust/

14. Hind, M., Wei, D., Campbell, M., Codella, N.C., Dhurandhar, A., Mojsilovic, A., Ramamurthy, K.N., & Varshney, K.R. (2018). TED: Teaching AI to Explain its Decisions. *AIES '19.* doi: 10.1145/3306618.3314273

15. Hoffman, R.R., Mueller, S.T., Klein, G., & Litman, J. (2018). Metrics for Explainable AI: Challenges and Prospects. ArXiv, abs/1812.04608.

16. IBM. (n.d.). Design for AI. Retrieved from https://www.ibm.com/design/ai/

17. Kessler, T.A., Stowers, K., Brill, J.C., & Hancock, P.A. (2017). Comparisons of Human-Human Trust with Other Forms of Human-Technology Trust.

18. Kizilcec, R.F. (2016). How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.*

19. Kocielnik, R., Amershi, S., & Bennett, P.N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. *CHI '19.*

20. Kulesza, T., Stumpf, S., Burnett, M.M., Yang, S., Kwan, I., & Wong, W. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing, 3-10.*

21. Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40,* 153-184.

22. Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors, 46*(1), 50–80. doi: 10.1518/hfes.46.1.50_30392

23. Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems - CHI 20.* doi: 10.1080/14639220500337708

24. Lim, B.Y., & Dey, A.K. (2009). Assessing demand for intelligibility in context-aware applications. *Proceedings of the 11th international conference on Ubiquitous computing.*

25. Lim, B.Y., Dey, A.K., & Avrahami, D. (2009). Why and why not explanations improve the intelligibility of context-aware intelligent systems. *CHI.* doi: 10.1145/1518701.1519023

26. Lyons, J.B., & Stokes, C.K. (2012). Human-Human Reliance in the Context of Automation. Human factors, 54 1, 112-21. doi: 10.1177/0018720811427034

27. Madhavan, P. & Wiegmann, D. A. (2007) Similarities and differences between human-human and human-automation trust: an integrative review. *In Theoretical Issues in Ergonomics Science, 8:4, 277-301.* doi: 10.1080/14639220500337708s

28. Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 53*(4), 356–370. doi: 10.1177/0018720811411912

29. Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *ArXiv, abs/1706.07269.*

30. Parasuraman, R., & Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors, 39*(2), 230–253. doi: 10.1518/001872097778543886

31. Parasuraman, R., Sheridan, T.B., & Wickens, C.D. (2000). A model for types and levels of human interaction with automation. *IEEE transactions on systems, man, and cybernetics. Part A, Systems and humans : a publication of the IEEE Systems, Man, and Cybernetics Society, 30* 3, 286-97. doi: 10.1109/3468.844354

32. Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *KDD '16.*

33. Selbst, A. D., Powles, J. (2017). Meaningful information and the right to explanation, International Data Privacy Law, Volume 7, Issue 4, 233–242. doi:10.1093/idpl/ipx022

34. Sokol, K., & Flach, P.A. (2020). Explainability fact sheets: a framework for systematic assessment of explainable approaches. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency.*

35. Springer, A., & Whittaker, S. (2018). What Are You Hiding? Algorithmic Transparency and User Perceptions. *ArXiv, abs/1812.03220.*

36. Stowers, K., Kasdaglis, N., Rupp, M., Chen, J., Barber, D., & Barnes, M. (2017). Insights into human-agent teaming: Intelligent agent transparency and uncertainty. *Advances in Human Factors in Robots and Unmanned Systems.* 149-160.

37. Wang, D., Yang, Q., Abdul, A., & Lim, B.Y. (2019). Designing Theory-Driven User-Centric Explainable AI. *CHI '19.*

38. Wolf, C.T. (2019). Explainability scenarios: towards scenario-based XAI design. *Proceedings of the 24th International Conference on Intelligent User Interfaces.*

39. Zhang, Y., Liao, Q.V., & Bellamy, R.K. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

40. Zhu, J., Liapis, A., Risi, S., Bidarra, R., & Youngblood, G.M. (2018). Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. 2018 IEEE Conference on Computational Intelligence and Games (CIG), 1-8.