On the Importance of User Backgrounds and Impressions: Lessons Learned from Interactive AI Applications

MAHSAN NOURANI, University of Florida,
CHIRADEEP ROY, University of Texas in Dallas,
JEREMY E. BLOCK, University of Florida,
DONALD R. HONEYCUTT, University of Florida,
TAHRIMA RAHMAN, University of Texas in Dallas,
ERIC D. RAGAN, University of Florida,
VIBHAV GOGATE, University of Texas in Dallas,

While EXplainable Artificial Intelligence (XAI) approaches aim to improve human-AI collaborative decision-making by improving model transparency and mental model formations, experiential factors associated with human users can cause challenges in ways system designers do not anticipate. In this paper, we first showcase a user study on how anchoring bias can potentially affect mental model formations when users initially interact with an intelligent system and the role of explanations in addressing this bias. Using a video activity recognition tool in cooking domain, we asked participants to verify whether a set of kitchen policies are being followed, with each policy focusing on a weakness or a strength. We controlled the order of the policies and the presence of explanations to test our hypotheses. Our main finding shows that those who observed system strengths early-on were more prone to automation bias and made significantly more errors due to positive first impressions of the system, while they built a more accurate mental model of the system competencies. On the other hand, those who encountered weaknesses earlier made significantly fewer errors since they tended to rely more on themselves, while they also underestimated model competencies due to having a more negative first impression of the model. Motivated by these findings and similar existing work, we formalize and present a conceptual model of user's past experiences that examine the relations between user's backgrounds, experiences, and human factors in XAI systems based on usage time. Our work presents strong findings and implications, aiming to raise the awareness of AI designers towards biases associated with user impressions and backgrounds.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI**; **User studies**; • **Computing methodologies** → *Neural networks*.

Additional Key Words and Phrases: Explainable AI, Cognitive Biases, HCI, User Studies, Conceptual Models

1 INTRODUCTION

Over the past decade, machine learning and artificial intelligence algorithms have been incorporated in different contexts and domains to make systems more intelligent and autonomous. Unfortunately, many of these so-called

Authors' addresses: Mahsan Nourani, University of Florida, Gainesville, Florida, , mahsannourani@ufl.edu; Chiradeep Roy, University of Texas in Dallas, Dallas, Texas, , cxr161630@utdallas.edu; Jeremy E. Block, University of Florida, Gainesville, Florida, , j.block@ufl.edu; Donald R. Honeycutt, University of Florida, Gainesville, Florida, , dhoneycutt@ufl.edu; Tahrima Rahman, University of Texas in Dallas, Dallas, Texas, , tahrima.rahman@utdallas.edu; Eric D. Ragan, University of Florida, Gainesville, Florida, , eragan@ufl.edu; Vibhav Gogate, University of Texas in Dallas, Dallas, Texas, , vibhav.gogate@utdallas.edu.

2160-6455/2022/1-ART1 $15.00
https://doi.org/10.1145/3531066

blackbox algorithms are hard to understand for the users due to the complexity of their inner logic [78]. This lack of transparency can cause users to experience problems due to an inappropriate mapping between their mental model of how the model works and the reality of how it works, which can lead to other problems such as over or under-reliance on the intelligent system [7].

To help solve these problems, researchers and practitioners have introduced eXplainable Artificial Intelligence (XAI) models, where the systems attempt to explain their decision-making process to the users [40]. Explanations can be anything from general information about and extracted from the model (e.g., post-hoc explanations [34]) to annotation of the input to highlight the features used in the decision-making process (e.g., [56]). For simplicity, in the context of this paper, we refer to instance-level post-hoc explanations as *explanations* and use them to test our hypotheses and generalize our findings.

Theoretically, explanations should help users build a better mental model of an intelligent system [79]. However, in practice, as the models get more and more complex, it becomes harder to explain them in a manner that is beneficial to the users—as also suggested by previous work in psychology (e.g., [27]). One major problem is that with exploratory intelligent systems and tools, system designers have little to no control over *when* users encounter inaccurate and accurate predictions. As a result, the order of observing accurate vs. inaccurate predictions may introduce unintended biases in a user's mental model of the system. For example, previous research has shown that the order of encountering wrong predictions significantly affected a user's perception of accuracy [44]. However, there is little understanding of the interplay between the order of observing system weaknesses and the presence of explanations with respect to the user's mental model of the system.

In this study, we incorporate an explainable intelligent system (an online user interface tool powered by an explainable deep learning model) with an exploratory task to test how the order of observing system weaknesses and strengths can affect user's mental model of the system, and whether explanation presence can help improve these shaped mental models. The intelligent system we used was a video activity recognition tool (with cooking videos) where users could query the system to find certain actions and objects in the videos. The task was simple but exploratory: users were provided with a set of kitchen policies, and they had to determine which of the policies were being followed and which were not in a set of cooking videos. During the study session, the order of the policies was manipulated to influence when participants experienced correct and erroneous system outputs. We ran a 2x2 user study controlling both policy order and explanation presence. Our results showed that users with positive first impressions formed a better mental model of system strengths, though they also made more errors due to over-reliance on the model's answers to queries. However, users who encountered more model errors early formed negative first impressions that ultimately lead to a limited mental model and underestimation of system capabilities. Our results provides a novel contribution through an empirical user study aimed to help intelligent system designers to be aware of human cognitive biases (specifically, anchoring bias and first impressions) when using intelligent systems. The findings of this paper inspired us to take a deeper dive into how biases that are associated with user's backgrounds, impressions, and experiences affect user behaviours and the outcomes of human-AI collaborations. To further our contribution, we categorize human-centered outcomes and factors based on usage over time in an empirically-driven conceptual model of users' past experiences and draw recent work from the literature in human-centered AI to demonstrate examples from the concepts and categories in our model. Our work presents strong findings and implications, aiming to make intelligent system designers aware of biases associated with impression formations when designing interactive, intelligent systems.

2 RELATED WORK

2.1 Mental Models in Explainable AI Research

Researchers in the human-computer interaction (HCI) community have been studying XAI systems from different angles. Explainability is central to the community efforts for developing responsible and fair AI models [48]. As

evidenced in Shneiderman's discussion on the role of Human-Centered Artificial Intelligence (HCAI) [65], the need for clear descriptions of how an intelligent agent makes its predictions/decisions is critical to establishing reliability, safety, and trust with such tools. With an emphasis on interactivity, Shneiderman argues that AI systems should not only provide post-hoc explanations—such as saliency maps or decision trees—but also design exploratory interfaces that allow users to develop a working understanding through investigative interactions. In other words, he argues users are capable of understanding and building better mental models of ML/AI systems when they are allowed to explore interactive, explainable models. Explainability and its impacts on humans behaviours have been examined through various perspectives, such as scope [2, 3], type [32], and the target users [12, 44, 62, 76]. One of the main focuses in XAI is to improve user understanding, by facilitating the construction of a mental model with an accurate picture of a model's limitations [25, 40].

User mental models of machine learning and AI algorithms can be defined as the representation through which users understand the system [40]. In other words, mental models reflect users' understanding of how a machine works and how their actions can affect the outcomes [4]. As they directly influence user perceptions of intelligent systems, it is important to study user mental models. Flawed mental models (whether inaccurate, incomplete, or based on false assumptions) lead to a variety of other problems, such as unmet expectations, frustrations [4], over-/under-reliance, and problems establishing and calibrating trust. With the involvement of AI/ML technologies in various aspect of human lives, it becomes crucial to study user mental models to build a general, scientific understanding of these concepts and what behaviours they can lead to and how they affect usage.

Mental models have been studied in different fields for a long time. For example, psychology researchers have studied how humans form mental models of the world. Human-centered AI research can benefit and apply the findings from psychology to improve the community's understanding of mental models of AI systems and incorporate techniques that are beneficial to users when they form them. For example, Hoffman et al. [22] develop and present a conceptual model of the process of explaining and discuss the role of mental models in the XAI pipeline through the lens of psychology.

Explainability and mental models are commonly linked together. Not only because improving user mental models is one of the main goals of explainability [40], but also because both influence other human factors with AI applications and tools. Many researchers have studied the reflection of mental models on user trust. For example, Holliday, Wilson, and Stumpf [25] discover that explanations can affect how users calibrate their trust in a model based on whether they were provided with explanations or not. They find that user *perceptions* of the model directly affected their abilities to trust, and without explanations, they were unable to develop a mental model of how the model works and therefore, gradually lost their trust as they continued working with the model. Kulesza et al. [**?**] also studied how certain features of explanations can affect mental model formations and trust. Based on their results, the relationship between trust and mental models of an XAI tool vary based on the soundness and completeness of explanations. For example, with highly sound and complete explanations, users build more solid mental models and highest trust in explanations. However, when explanations are too detailed, users show improvements in some aspects of the task, while their trust in explanations decrease. Examples like these warrant for studying and exploring mental models in intelligent explainable tools.

Researchers in the HCI community focus on implementing and discussing various interactive, visual techniques that can improve user mental models and understanding of a machine learning system. Some of these techniques provide an instance-level explanation for a specific input (e.g., [51, 57, 63])—that is also referred to as local explanations—while others expose the model's inner workings without necessarily focusing on specific examples/outputs (e.g., [23, 47])—i.e., providing global overviews (or explanations) of how a model works. More recent work has gravitated to the use of global explanations by using ontological maps [9] or layered graphs [24] to communicate how model parameters or input features may influence one another to facilitate more efficient mental model construction of the global system's performance via post-hoc explainability. These techniques

attempt to provide a general impression of how the system performs by summarizing patterns of behavior or collating trends in system responses over the data set. On the other hand, some methods were established in an attempt to balance the cognitive load required for local explanations without sacrificing model-explanation fidelity. For example, Abdul et. al. [1] designed a hybrid approach that utilizes bar charts to express the relative importance of different features and incorporates small multiples to show how features influence regressive models for the entire dataset. In turn, they show how their hybrid method is preferred by users and optimizes the communication of model performance in a human review task. In different words, they showed that their proposed approach decreases a user's cognitive load without decreasing the accuracy.

To be able to build a solid mental model of a system, users need to be exposed to both its strengths and weaknesses. While global interpretability can, in theory, help improve user's understanding of model shortcomings and capabilities, they are more challenging to generate in practice [2]. Moreover, many stakeholders do not necessarily benefit from global explanations as they would with local explanations. Wortman Vaughan and Wallach [76] describe two user mental models, *structural* and *functional*, and caution system designers and researchers on making a distinction between the two. The former allows the stakeholders to understand how a system works (could be achieved via global explanations) while the latter helps them to use the system without necessarily understanding how the model works (could be achieved via local explanations). Which of these mental models is sufficient for each stakeholder remains an open challenge, and might be relevant to other factors, such as user task and their level of corresponding expertise to the task. Generally speaking, the majority of work in the HCI, ML, and AI communities has been focused on instance-level explanations that justify model predictions per output [2] and depend on the user's functional mental models. In this paper, we will also focus on local explanations, providing system responses for individual frames and corresponding queries.

2.2 Measuring Mental Model

As interactive interfaces and data visualization techniques may be utilized to enhance user mental models of the limitations and competencies of intelligent systems, to effectively test the effectiveness of the given techniques, researchers need to measure and capture user mental models. Previous research demonstrates that it is not easy to measure mental models due to their temporal nature and their influence on user disposition [42]. Since their initial description in 1943 [10], mental models are generally inferred from a variety of user study techniques, such as think-aloud approaches [61], interviews [37], and well-constrained survey questionnaires [20, 22]. Research on mental models in intelligent systems shows that as users work with an intelligent system, they develop more robust mental models, thus relying less on their dispositional trust and more so on their experiential trust [35, 36]. In XAI communities, different people have reviewed and proposed different techniques and measures to quantify and qualify a user's mental model of the algorithm (e.g., [22, 40]). Prediction tasks are one of the commonly used techniques in this field, which require subjects to estimate model predictions and performance *after* they have been exposed to its outcomes [21, 22, 62]. Given a novel sample, users are asked to predict and estimate how they think the model will respond; with controlled choices, the unique differences between the options serve as a proxy for what users believe about the system. In this realm, Poursabzi-Sangdeh et al. [52] found that simpler models with fewer features enable users to predict and simulate the model predictions. As reflected in cognitive science, when more models are required to make an inference, the more challenging it is for individuals to understand the complexity of the problem [27]; therefore, the emphasis is on making visualizations that summarize the autonomous system in a tractable way to assist in the valid construction of mental models.

2.3 Cognitive Biases in XAI: First Impressions and Overconfidence

Mental models tend to be simplified heuristics used as foundations for more complicated thought; however, assuming that they are free of fault and being affected by cognitive biases is a mistake hinted at by HCI and Psychology researchers extensively.

In his book, Baron [6] lists and classifies more than 50 different known and discovered cognitive biases. One of these classes is *motivated bias*: Humans have beliefs that are aligned with the truth and can serve as a basis for decision-making. That is to say that humans are motivated to use their beliefs to evaluate new information and then adjust their beliefs based on the veracity of the information they receive, as a consequence of not wishing to be seen as incorrect. Of the biases in this category, the present work explores the *primacy effect*–also studied under different names (e.g., anchoring bias [8, 64, 77], order bias [55], and first impressions [30, 44]).

The primacy heuristic refers to when people make assumptions about someone/something earlier in their encounters and are anchored towards those assumptions later on. This overlaps with the concept of *confirmation bias*, where people tend to collect redundant information that aligns with an initial assumption as opposed to contrasting information that can lead to a more complete understanding by refuting their assumption. Lighthall and Vazquez-Guillamet [33] argue two causes for this bias: (1) [in]correct assumptions (where a person's decision on some variable is biased by another variable); and (2) a psychological tendency to rely on a[n incorrect] decision they already made rather than restarting their decision-making process. With more people encountering and using ML/AI systems each day, it is important to study how people form first impressions of these algorithms.

First impressions are explored in various human-centered fields, such as psychology [53], social technology (e.g., first impressions in news [11] or social media), social sciences [5], and decision-making [77]. First impressions are also extensively explored in Human-Robot Interaction (HRI)—where researchers study how human's first impressions of a robot can influence their trust in and communication with the AI system and feelings of uneasiness like those associated with the uncanny valley [19, 49, 50, 80]. However, in the recent explorations of human factors in explainable intelligent systems (i.e., human-centered ML/XAI), few discuss the consequences of potent first impressions. In this paper, we tend to showcase some example work in this domain and lay the groundwork for future researchers to focus on unexplored challenges within this realm or propose approaches to mitigate this bias. Myriad factors influence how first impressions of ML/AI systems develop, from the user's emotional and cognitive state to the physical conditions of the interaction. In recent work, Kim et al. [29] show and discuss that *when* a model makes an error can strongly influence user reliance. They found that if users experience the errors earlier, their reliance decreases, while experiencing errors later-on can only influence their reliance temporarily. In a prior work (Nourani, King, and Ragan[44]), we studied how domain expertise affects users' first impressions of an intelligent system and how these impressions impact their trust and its evolution over time; we found significant differences in impression formations based on domain expertise, which we will discuss in more detail in Section **??**.

Beyond the formation of first impressions, there are also concerns about the development of rigid heuristic beliefs that lead to decision support tool neglect while constructing a mental model of a decision aid's performance. This abandonment of system support is sometimes referred to as overconfidence. Siek and Arkes [66] examine how overconfidence develops in human-AI paired decision-making. In a series of experiments, they asked participants to predict jurors' opinions of assisted suicide given a handful of attributes (e.g., age, political party affiliation, and alcohol consumption). In most conditions of their experiment, participants were provided with system predictions from a regression model. Participants were told that the model was accurate 77% of the time and, even though this was controlled to not always be the case, they found that participants generally favored their own "gut feeling" over the model's advice. This reliance on oneself over a model shows how overconfidence can develop and they discuss how challenging it can be to correctly redress the associated biased behaviors. Their result

further emphasizes the challenge of controlling for these unintentional biases and the caution one must take as they develop intelligent tools and training.

Finally, as seen in Van der Waa et. al.'s [75] work, additional time pressure and a delay in the resulting consequences of human-machine paired decisions in a high-stakes simulated medical triage scenario further emphasizes the requirement for explanations to not only be available to the human reviewer but also supply ample time to fully comprehend and take responsibility for the AI's behavior. In our task, we remove the time pressure to prevent the additional demands and stress participants may experience and also provide various explanations at once to give participants the freedom to use the tools they find most applicable. Our work, focused on the development of a user's mental model in a more *exploratory* scenario, deviates from much of these past experiments because we have less control over what a user observes and how they use the system. This makes our work closer to decision-making tasks in more realistic settings.

3 EXPERIMENT

We conducted a human evaluation to understand how first impressions of intelligent systems can influence user mental models, as well as task performance and reliance on the tool. We also sought to learn whether explanations can help bypass the biases formed in the earlier encounters with model predictions. In this section, we describe our experiment design in more detail.

3.1 Explainable System

3.1.1 System Context. For this study, we sought an open-ended scenario where users could explore the system and build a mental model of how it works. With some intelligent systems, errors can be tolerated to some extent and they may not be fatal. That is why it might seem unnecessary for the users to build mental models of the system. However, some systems naturally require a human agent to monitor the outcomes and predictions rather than automatically accepting failures without worrying about the consequences. Examples of such systems, and our system of choice, include video activity recognition systems, where a model can be trained to automatically detect activities that take place in the videos. In real-world scenarios, activity recognition has many use-cases and can be critical due to physical limitations and time constraints. Some examples include fire detection [31], airport security [74], smart hospitals [71, 82], and elderly care [28]. Since we desired a task where users are novices and do not require any certain expertise or professional training, we chose a cooking video scenario where the system was designed to identify cooking-related tasks in a kitchen. In the rest of this chapter, we briefly describe the model and interface we used for the system we designed for our experiment.

3.1.2 XAI Model. The XAI model used in this study was trained on a pre-annotated dataset of cooking videos called the TACoS dataset [58]. Note that the development of the XAI model is not a part of the contributions presented in this paper, as the model was only used to serve the goals of the experiment while using a real explainable model for the system. More details on the specifics of the model can be found in our previous work [59]. Here, we provide an overview of the model to help readers understand the basis for the model capabilities and explanations.

In the TACoS cooking videos [58], each frame of each video had a set of labels (which we call ground labels) that summarized the activity taking place in the video (for example, *{"wash", "carrot", "sink"}* in frames where a carrot was being washed in the sink). The problem was formulated as a multi-label classification problem where given each frame of the video, the model had to assign the correct labels to it. Each label was modeled as a binary random variable where 0 and 1 indicated that the label was off or on respectively. We implemented a two-layer architecture where the first layer comprised a deep neural network based on GoogleNet [69] that converted each frame into a set of noisy labels and the second layer used a dynamic version of a tractable probabilistic model called a cutset network [54] that modeled a conditional probability distribution of the ground (true) labels given

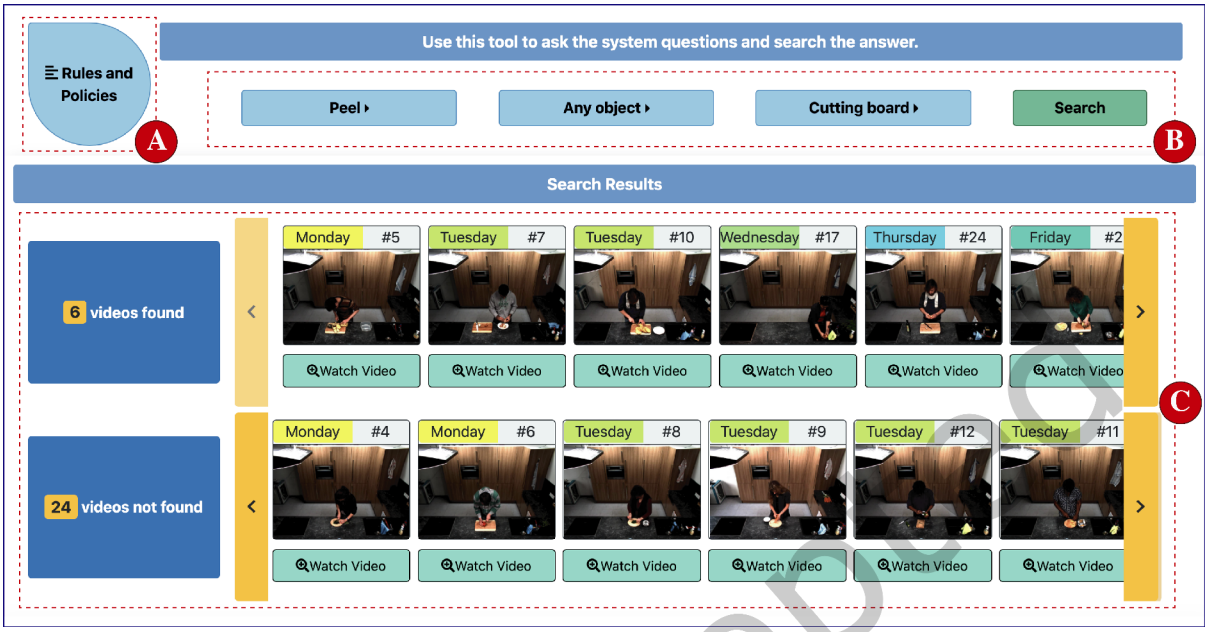

Fig. 1. The main overview of the user interface. By clicking in the top left corner (A), a panel opens from the left side of the screen that includes a list of policies. Here, users recorded the kitchen's compliance with each statement. (B) Users selected components from three drop-downs to build a query and search for it among the videos. (C) The search sorted the thumbnails into two categories: matching and non-matching videos. By showing a thumbnail preview of each video, their assigned unique ID, and their corresponding weekday, users could select *watch video* to inspect and explore more.

the noisy labels from the neural network, i.e., $P(G^{1:t}|E^{1:t})$ where $G^t = \{G_1^t, ..., G_n^t\}$ is the set of ground labels at frame t and $E^t = \{E_1^t, ..., E_n^t\}$ is the set of corresponding noisy evidence labels. The top layer was designed as an "explanation" layer in order to (1) remove the noise from the GoogleNet labels and (2) model the temporal relationships between the ground (true) labels. The model was trained on 30 videos with a vocabulary of 35 labels. Explanations were computed on the final trained model by formulating them as two standard probabilistic inference queries: posterior marginal (MAR) and top-k most probable explanation (MPE). The MAR query seeks to estimate the probability of the true label given noisy labels obtained from GoogleNet while the top-k MPE query seeks to find the top k most likely assignments to the true labels.

3.1.3 Main Interface. We designed an interactive video activity searching tool to allow users to build specific queries and sort the videos from the dataset. The design of this tool was motivated from and was built on findings from an earlier version of this tool [46] that focused on a more controlled scenario in which participants would see a set of specific examples. By grounding the exploration capabilities of the design, we tailored a simpler interface for a similar model where participants were asked to review and evaluate the correctness of the model predictions to yes/no queries before we evaluated their mental models of the XAI system. While the user experiment led to insightful findings, it failed to capture differences with user mental models, to which we believe a few factors contributed. Most importantly, to form mental models, users need to explore the interface freely and on their own terms and pace (as they most likely would in the real-word applications). These observations and findings from our prior work inspired us to opt for a new exploratory, open-ended interface and system design with a less controlled task to provide more opportunities for users before we try to capture their mental models. In

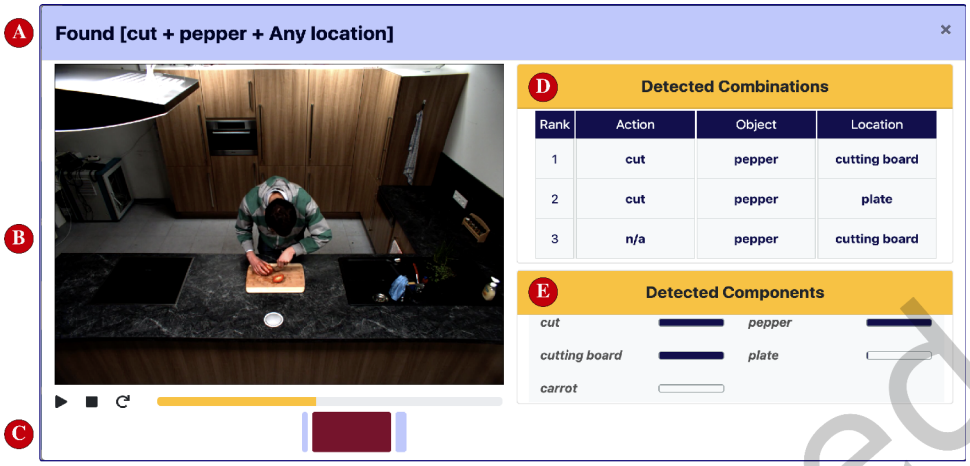

Fig. 2. When clicking on the *watch video* button in the main interface, as seen in Fig. 1, participants would see a modal to allow them to watch the video. (A) showed the selected query and whether the query is found or not found in the video (B). If they were in the *explanation presence*, they were shown all the video segments that were used to come up with the answer (found/not found) under the progress bar (C). They were able to click on each of the available segments to see the model justification based on the relevant activities found in the segment (D), as well as the system's confidence score in all the components it detected within the selected segment (E).

our new tool, we define each activity using three component types: *Action*, *Object*, and *Location*. Fig. 1 shows the overview of the interface. The top of the screen has a simple query builder where users can input specific component combinations or select a generic form (e.g., any action). After searching, the interface would organize the videos into two lists based on whether the model found the searched activity in each video or not. The XAI system showed thumbnails for each video to distinguish them from the other videos in the list. Each video was assigned an id number and day of the week to help users track how the system responded.

3.1.4 Explanation Interface. By clicking on a thumbnail, a modal overlay would open where users could watch the video and see the model explanations to examine why the video was categorized as a match (or non-match) for the query. Fig. 2 shows the three explanation elements for each video that aimed to assist the users in understanding why the model matched the query with the video. Directly under the video progress-bar (Fig. 2.C) was a series of *video segments* that highlighted the most relevant set of frames used by the model to answer the current query. Clicking a video segment updated the information presented in the other two explanation elements: (1) The *detected combinations* (Fig. 2.D) listed the top 3 queries that the model associated with the currently-selected video segment and (2) the *detected components* (Fig. 2.E) showed the model's confidence about the activity components detected separately in this video segment. Note that the explanation interface was similar to the one used in our prior work [46], showing high-fidelity, post-hoc explanations from the XAI model.

3.2 Research Goals and Hypotheses

For this study, we were primarily motivated to understand the role of first impressions on a user's mental model formation. As one of the main motivations behind XAI research is to improve user understanding and mental models of intelligent systems [18], we deemed to test whether and how the addition of explanations can affect user mental models, given that users might have formed initial biases in their assumptions towards the system. Therefore, we designed a policy-verification task, where the system described in Section 3.1 was used to verify

whether a set of kitchen guidelines and policies are being followed by the people performing cooking activities. This was a task, exploratory enough to allow users to freely test and observe various system predictions to build a mental model of both system weaknesses and strengths. Moreover, with an open-ended and real-world scenario, we are able to generalize our findings to other intelligent decision aids. We designed a study where participants observed the same set of policies, while we controlled that earlier in the usage, some observed policies that expose system weaknesses while others observed the policies that exposed system competencies. Also, with each order, some participants were provided explanations while others were not. By comparing these conditions, our evaluation explored how users' interpretation of the *same* system may be different based on their experience of system performance with or without the addition of explanations. These goals and research question are summarized in the following set of hypotheses:

- **H1**: Encountering model weaknesses early-on will lead to less usage and reliance compared to encountering model strengths early.
- **H2**: Positive first impressions can improve user mental models while negative first impressions can impair them.
- **H3**: Regardless of the order of encountering model weaknesses and strengths, model explanations help decrease or eliminate the effect of anchoring bias on user reliance on the system.
- **H4** The addition of explanations will significantly improve user task-performance and mental models by increasing their understanding of AI system weaknesses and competencies.

3.3 Experimental Design

After describing the intelligent system and the goals of the study, we turn our attention to the study design details.

3.3.1 User task. Using the XAI system described in Section 3.1, we sought an exploratory task to allow the participants to use and experience the system and build a mental model of it. As we were also considering a task that did not require any expertise or professional training, we used a kitchen policy scenario, where participants were given a set of kitchen rules and policies and were asked to determine, using the system, which of the policies were being followed by the kitchen staff.

We generated intricate policies that generally required users to build and test multiple queries in order to encourage further use of the intelligent system. Each policy was designed to either expose *model weaknesses* (i.e., components that were misidentified or remained unidentified) or *model strengths* (i.e., those components known to be consistently identified correctly). Due to this design, we ended up with 4 policies focused on system weaknesses and 4 policies focused on system strengths. Additionally, we used one policy as attention check, which was unique since it was not ubiquitously followed by the kitchen staff, but would sound logical to users not watching the videos: "Employees wash their hands immediately after entering the kitchen". Ultimately, participants received nine policies to interpret and were asked to determine their truthfulness in a set of thirty cooking videos. Policies were simple statements of fact that used components available in the query builder, like "Employees must not use *pineapples* more than 3 days a week" or "*Carrots* are only *cut* on rectangular cutting boards". Additionally, since the post-task questionnaire asked users to report on their mental models and usage of the system, we repeated components in multiple policies to increase memorability and to support user understanding.

The interface included a list of policies (a hidden panel on the left side of the screen until the participants decided to open them by pressing the "Rules and Policies" on the top left corner of the screen, as seen in Fig. 1.A), and participants indicated if each was met with yes and no buttons.

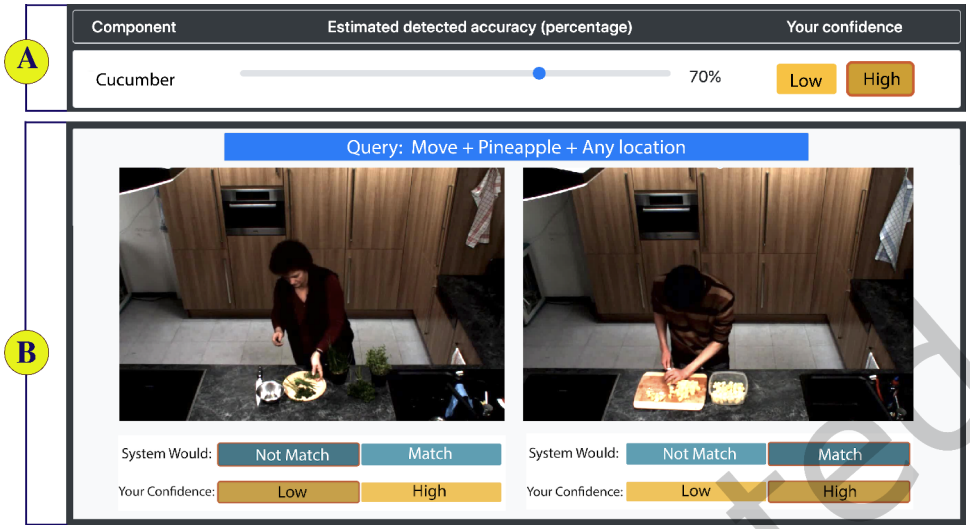

Fig. 3. Examples of the mental model questions for the user study. (A) The user estimated the accuracy for cucumber was 70% and had a *high* confidence in their estimation. (B) Frame-query estimation where the user guessed whether the system matched each frame to the query and rated their confidence in their response.

3.3.2 Conditions. To address our goals and hypotheses, we designed a 2x2 between-subjects user study with two independent variables: (1) *policy order* and (2) *explanation presence*. Participants were assigned one of the four conditions randomly and everyone completed the same task. We controlled the order of observing policies so that some participants were exposed to system weaknesses first while others were exposed to system strengths first. We also maintained that the attention check policy would always remain in the middle of the list of policies. Ultimately, all participants observed the same set of policies, but with varying order. In pilot testing, we observed that participants consistently examined each policy in sequence starting from the top of the list, so we relied on this behavior to control for the policy order factor. We also updated the system interface described in sections 3.1.3 and 3.1.4 to match the assigned condition. We changed the video thumbnails to show the most relevant frame for the *with explanations* conditions and the middle frame for the *no explanations* conditions. Also, while those in the *with explanations* conditions observed all the three explanation elements within the explanation interface, the participants in the *no explanations* conditions were only provided with the video player (i.e., only elements (A) and (B) in Fig. 2).

3.3.3 Measures. In addition to interaction logs, we asked participants to complete four post-task questionnaires designed to quantify and explore the limits of users' perception of the system's strengths and weaknesses (i.e., their mental models), as well as usage and reliance. We selected two types of questions for assessing mental models. The first, as shown in Fig. 3.A, asked users to estimate the detection accuracy for eight activity components we selected that appeared in the policies frequently. Some of these components were from model weaknesses (e.g., *pineapple*) and some of them were from model strengths (e.g., *carrot*). Estimation of accuracy is an established known method for estimating general user understanding of model performance and mental model of system capability (e.g., [26, 40, 44]). With a slider, users indicated how accurately the system detected each component (0–100%) and also marked their confidence (low or high) in their answer. In the second question, as seen in Fig. 3.B, the participants were given an activity query with a set of 4 video thumbnails and were asked to *predict* whether the system would categorize each thumbnail as a *match* or *non match* using their mental model of the

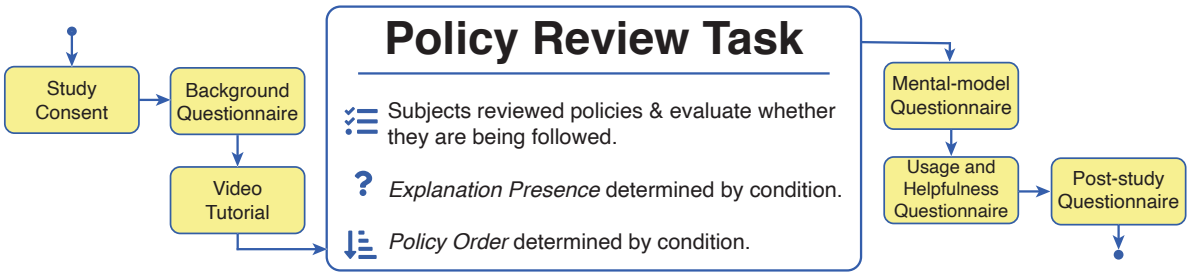

Fig. 4. An overview of the study procedure. Participants were mainly asked to review a set of kitchen policies and verify whether each of them were being followed. After finishing the main task, we measured their mental model of the model's strengths and weaknesses through two sets of prediction tasks (as seen in Fig. 3).

system. They were also asked to rate their confidence in their prediction (low or high). We provided three queries, each with four assigned thumbnails, making a total of 12 frame-query predictions per participant. This measure was inspired by *prediction tasks* which are another established method in assessing and measuring the user's mental model of AI/XAI systems [22, 40].

We then asked the participants to rate both usage and helpfulness for each interface element on a 5-point Likert scale. These measures were adjusted for participants based on their explanation condition (i.e., they were only asked about components they saw). Finally, they rated their estimation of the model's overall accuracy in percentage, as well as answering a few free-response questions describing any noticeable weaknesses or feedback to the researchers.

3.4 Procedure

In a single online session, participants completed the following, as summarized visually in Fig. 4. The research was approved by the organization's institutional review board (IRB). All participants took about 20 minutes to verify all the policies. After observing the study's informed consent, participants were asked to complete a brief demographic background questionnaire.

Participants were then introduced to their task via video tutorial that described the task as well as how to form a query by providing an example. To help participants understand the task better, we designed a tutorial video, introducing a hypothetical restaurant owner who asks the participants to use the intelligent tool and verify whether the kitchen rules are being followed by her employees by inspecting the surveillance footage from the past week. Participants were informed that one food was prepared by one chef per video and that there were six videos per day of the week (i.e., 30 videos in total). The tutorial then described how to use the tool and how the task can be achieved. To avoid learning effects, the tutorial used an extra policy to demonstrate the interface functions. We created two versions of the video for each of the *with explanations* and *no explanations* conditions. We also included a summary of the tasks and important considerations on the main page under the query building tool for users to refer to during the study.

After the tutorial, the main task had participants verify nine relevant kitchen policies listed in a sidebar. After answering all nine policies, the participant continued to the post-study questionnaire to evaluate their mental model and understanding of model weaknesses and strengths (more detail provided in Section 3.3.3).

3.5 Participants

We recruited a total of 116 participants from the university graduate and undergraduate students to complete the study online for class credit. The participants consisted of 78 males and 38 females. After carefully investigating the responses, we removed a total of 6 participants since they did not pass the attention check. Of the 110

remaining participants, 54 observed explanations: 28 of whom saw strong policies first and another 26 observed the weak policies first. Of those provided no explanations, 29 observed strong policies first while the remaining 27 initially saw weak policies. All participants were compensated, including those who did not pass the attention check.

4 RESULTS

In this section, we present the measures of our study and provide an analysis of the results. The findings of this paper were previously accepted and presented as a CHI 2020 extended abstract [43] and later, an IUI 2021 conference paper [45]. The current manuscript serves as an extension of our prior work, in which we provide a conceptual model of user past experiences in the lieu of the results from our previous findings and other relevant work. Thus, we draw conclusions from these findings to discuss the conceptual model user's past experiences. To learn more about the conceptual model, please refer to Section **??**.

Before performing data analysis, two steps were taken to avoid certain problems caused by performing an online study. To ensure the quality of participant responses without having a researcher present during the study sessions, we added an attention check policy and removed all of whom did not pass the test. Additionally, to account for some participants taking breaks during the task, we adjusted the task completion time by not counting any period of inactivity longer than five minutes. For each of our measures, we used a two-way factorial ANOVA for the main effect and Tukey HSD post-hoc testing for significant interaction effects, when applicable.

4.1 User-task Performance

First, to test our hypothesis about user-task performance, we tested both task time and task error to test. Task time is defined as the amount of active time spent on the policy review task. Task error was measured as the proportion of policies that the participant answered incorrectly. No significant effect was found for *explanation presence*. However, participants in the *weak first* conditions had significantly less error in their answers to the policy questions than participants in the *strong first* conditions, with $F(1, 106) = 6.55$, $p < 0.05$, $\eta_p^2 = 0.058$. No evidence of an interaction effect between *explanation presence* and *policy order* was observed. Additionally, no significant effects were observed on task time. Participants in the *with explanations* conditions made significantly fewer queries per policy compared to participants in the *no explanations* conditions, with $F(1, 110) = 4.30$, $p < 0.05$, $\eta_p^2 = 0.045$. Fig. 6.a shows the distribution of the task-error results across the conditions.

4.2 Component Accuracy

After completing the policy-review task, participants were asked to estimate the model's detection accuracy (percentage) for several components as described in Section 3.3.3. An example question for this measure is shown in Fig. 3.A. We selected these components so that five corresponded to system weaknesses (low model accuracy) and four to system strengths (high model accuracy). We compared the participants' perceived accuracy of each component with the system's actual accuracy for that component. Since our task and interface primarily had participants focusing on the matches returned by the system, we selected the system's positive predictive value of each component as the metric for system accuracy. Additionally, we only considered system performance on the videos that were used in the task.

For analysis purposes, we used the average error in percentage for both weaknesses and strengths for each participant separately, i.e., two metrics per participant. A similar approach was used for the confidence scores. The reason for this decision was to be able to compare the user's mental model of both system weaknesses and strengths and understand how each independent variable affected this understanding. We will discuss each of the two separately below:

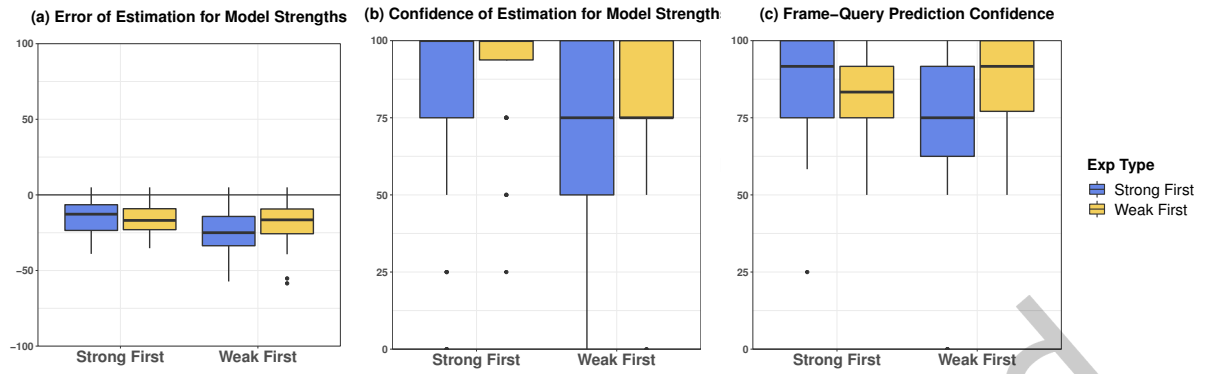

Fig. 5. Mental model metrics. (a) Participants' error of estimation for component accuracy (below 0 is underestimation). (b) Percentage of components for which participants rated as being confident in their estimation. (c) Percentage of frame-query pairs for which participants felt confident in their predictions. The last two plots are based on strength-detection (as described in Section 4.2)

Weakness Detection: For components that corresponded to system weaknesses, the statistical tests did not indicate significant differences across the conditions for neither the accuracy nor confidence.
Strength Detection: For components that corresponded to system strengths, participants who observed weaknesses first significantly underestimated the model's detection accuracy compared to those who saw strengths first, with $F(1, 106) = 6.24$, $p < 0.05$, $\eta_p^2 = 0.056$. Additionally, participants who observed weaknesses early-on were significantly less confident about their estimations compared to those who saw strengths early, with $F(1, 106) = 3.94$, $p < 0.05$, $\eta_p^2 = 0.036$. We did not observe any significant effect based on *explanation presence* on the user's strength-components' accuracy estimation or the confidence in their estimations. Fig. 5.a and 5.b show participant responses and their confidence across the conditions, respectively.

4.3 Frame-Query Prediction

Additionally, we asked participants to predict what output the system would have on a given frame-query pair, as observed in Section 3.3.3. An example of this prediction question can be seen in Fig. 3.B. We did not observe any significant differences among the conditions for the prediction accuracy. The mean prediction accuracy was $M = 0.599$ with a standard deviation of $SD = 0.127$ for participants with explanations and $M = 0.601$ with a standard deviation of $SD = 0.148$ for participants without explanations. This shows that users' estimations were barely better than guessing. However, a significant effect was observed on the confidence participants had in their responses. Participants with explanations were significantly more confident in their predictions than those without explanations, with $F(1, 106) = 4.12, p < 0.05, \eta_p^2 = 0.035$. There was also a significant interaction effect between explanation presence and policy order with $F(1, 106) = 5.20, p < 0.05, \eta_p^2 = 0.047$. A Tukey multiple comparison test showed the following significant interactions: Among the participants with no explanations, those who observed strong policies first were significantly more confident than their counterparts ($p < 0.05$). Participants with system explanations and strong policies first were more confident than those with no explanations and weak policies first ($p < 0.05$). Finally, of the participants who observed policies reflecting weaknesses early on, those who had system explanations were significantly more confident than those without explanations ($p < 0.01$). No difference in these effects was observed by splitting the frame-query pairs into those

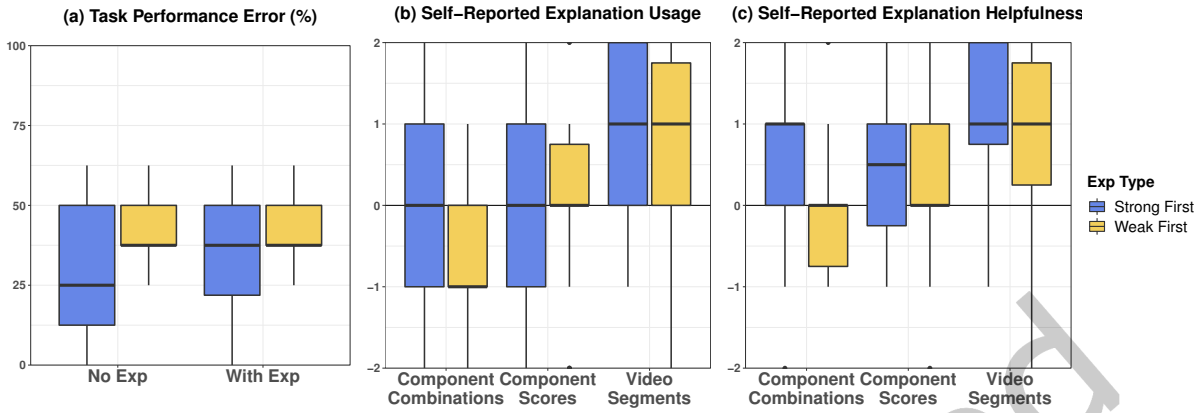

Fig. 6. Reliance and Usage metrics. (a) Participant error on the policy task (Percentage). (b) Responses to the question "How much did you use this element?". (c) Responses to the question "How helpful did you find this element?". The last two were measured on a 5-point Likert scale, with higher values indicating a higher rating of helpfulness and usage.

corresponding to system strengths and system weaknesses as we did with component accuracy. Fig. 5.c shows the confidence of the participant's responses among the conditions.

4.4 Explanation Usage and Helpfulness

After finishing the mental model questions, we asked the participants to report their usage of different interface components and how helpful they found them during their interaction period. Particularly, we were interested in the responses from those in the *with explanations* conditions about the provided system explanations; i.e., video segments (Fig. 2C), detected combinations (Fig. 2D), and detected components (Fig. 2E). Both usage and helpfulness were measured through a 5-point Likert scale. To run a more accurate analysis based on these three explanation types and *policy order*, we defined *explanation type* as a new independent variable for the analysis, and then performed a two-way independent ANOVA on explanation usage and explanation helpfulness. The results show participants who encountered weaknesses first reported a significantly lower rate of usage of system explanations than participants who encountered strengths first, with $F(1, 156) = 4.76, p < 0.05, \eta_p^2 = 0.030$. Additionally, we found that regardless of policy order, participants strongly preferred the video segments (Fig. 2C) in terms of both helpfulness and self-reported usage, with $F(2, 156) = 9.77, p < 0.001, \eta_p^2 = 0.111$ for explanation helpfulness and $F(2, 156) = 16.70, p < 0.001, \eta_p^2 = 0.176$ for self-reported explanation usage. We also analyzed user behavior—captured through interaction logs—to understand the usefulness of explanations by measuring how many queries participants performed on average for each policy. Participants who had system explanations completed the policy review task with significantly fewer queries per policy than participants who did not have system explanations, with $F(1, 106) = 4.94, p < 0.05, \eta_p^2 = 0.045$. No effect of policy order was observed for the number of queries made. Fig. 6 shows the self-reported usage and helpfulness of the different explanation types and the number of queries performed based on condition.

5 DISCUSSION

Our results demonstrate significant effects of first impressions on mental model formation, user reliance, and usage of the intelligent systems. In this section, we discuss the general indications of our results as well as their limitations and provide implications for system designers and opportunities for future work.

5.1 Interpretation of the Results

Participants in the *strong first* conditions had significantly more user-task error compared to those in *weak first* conditions. While this might seem counter-intuitive, it can be explained when compared to the findings from usage and helpfulness, as those who encountered system strengths earlier used explanations significantly more and found them to be significantly more helpful in the task compared to those who encountered weaknesses early. This indicates that observing strengths first can cause users to rely on the system more than they should (i.e., automation bias), while seeing weaknesses in the beginning can prevent this problem.

On the other hand, users in the *weak first* condition had problems forming their mental models of the system competencies and strengths. They significantly underestimated the system capabilities while also having less confidence in their estimations. These users are skeptical of system strengths but not confident in their skepticism because the weaknesses they observed earlier obscured their judgment of the system capabilities. This causes them to rely more on themselves rather than the model, leading to more confusion when shaping their mental model.

Participants in the *with explanations* conditions made fewer queries on average to answer each policy than those in the *no explanations* conditions, while task performance was similar regardless of the presence of explanations. One interpretation of this result is that the presence of explanations allowed users to achieve the same level of accuracy in fewer queries, thus they had increased task efficiency in terms of how many questions they needed to ask before being able to make their decision. However, this is only a hypothesis and due to the open-ended nature of this aspect of the study more research is necessary to help verify this effect.

We designed the frame-query prediction task to measure the user's granular mental model based on the specifics of the system. Though we did not observe any significant effects on the user's prediction, we did observe significant effects for the user's confidence in their prediction. Participants were more confident about their mental models when explanations were present. However, given that the mean for their original predictions were consistently around 50% in all the conditions (which is similar to guessing), we can conclude that these relatively high reported confidence scores are overconfidence. Our interaction effects show that without explanations, users in the *strong first* condition were more confident about their mental model, which we suspect is due to their automation bias, as discussed before. However, we observe that with explanations, users—regardless of their policy order—were more confident about their estimations compared to without explanation condition in *weak first* order. This might indicate that users can experience overconfidence in their mental model either when explanations are present or when strengths are observed earlier. However, we observed this overconfidence and overreliance through multiple tests for *strong first* order, showing that the order effect plays a more important role on a user's mental model than explanation presence (this can be supported by our results related to user-task error: users in *weak first* condition made fewer errors regardless of their explanation condition). This suggests that explanations alone cannot solve the strong bias created by first impressions.

Overall, these results suggest that unlike the general belief that model explanations can increase user understanding, they might not necessarily be beneficial. Explanations might cause a misconception in the users that they understand how the model works when, in fact, they do not. As shown by previous research in psychology, overconfidence (in this case, in the form of overprecision) can have serious consequences [14, 41]. Similarly, previous research suggests overreliance can cause several problems [7, 62], and our results provide a clear example of users making more errors due to automation bias. First impressions have strong influences on human's minds

towards information [72], and as shown by our results, they can be strong against automated systems as well. We would encourage future research into mitigating such biases, as they can have lasting effects on users' minds. More intensive and meaningful user studies are needed with realistic systems—as other researchers (e.g., [2, 12, 38, 39]) have also argued—to expose such biases and find techniques to (1) make users aware of their biases, (2) prevent users from forming new biases, and (3) help users rectify their own misconceptions and inaccuracies in mental models.

5.2 Design Implications for Intelligent System

With more complex and exploratory systems, the role of instructions and guided training becomes more inevitable; that is, allowing the users to use the system without interventions might affect how their mental models are shaped. With more critical tasks, it might be beneficial for the system to guide the formation of the mental model early-on to help users develop a more accurate foundational understanding of the system before actually using it in practice to make important decisions. Through this initial training phase, designers can control what kind of predictions users observe and in what order they are observing them. These decisions are task-dependent and can be made based on the priorities in that system. For instance, if sacrificing human-task accuracy (due to errors made from automation bias) to encourage the formation of more accurate mental models is acceptable, the introduction might focus more on showing system strengths earlier in the usage. Designers might also choose to sacrifice the mental model formation since they want to limit the number of mistakes made by the users, and thus, they can focus on highlighting more errors earlier in the usage. However, most designers might strive for the best of both worlds: limit the user mistakes by avoiding automation bias while allowing users to maintain an appropriate mental model of the system. Based on our findings, users who observed strengths earlier made more errors but formed a better mental model of the system strengths. Considering this finding, in the initial training, designers can guide users' early observations toward model strengths but also intervene and show errors occasionally to balance users' attention with errors as well. When errors are shown, designers can focus more on explaining why they happen. This can be done by altering explanation type, scope, and focus and differentiating it from the explanations provided for the correct predictions. Note that this is only possible in guided training as designers know what instances are correct and which are wrong.

Theoretically, a higher-level explanation could help users scaffold more accurate mental models by first introducing how the system works before using the instance-level explanations. Previous research suggests that global visualization and explanations can help users form a more appropriate perception of how the model works [40]. Allowing users to explore and understand how the model works on a higher level might help users form a mental model before encountering the intelligent system for the first time. Future research needs to test the extent of information sufficient for global visualizations for mental model formation, and whether this approach is effective for avoiding ordering and anchoring biases when using instance-level models. Finally, designers need to consider the effect of first impressions when designing explainable interfaces and be aware that the sole addition of explanations cannot circumvent bias formation. Comparing various types of explanations against one another (e.g., *why* and *how* explanations [2, 12, 32]) to understand which method works better against certain biases, or incorporating multiple explanation scopes within one interface might allow users to decide what they want to explore to understand the model decisions better. For example, with an analytical tool, a user can look for different types of information and explanations from the model when encountering errors to improve their understanding of the model.

5.3 Limitations and Opportunities for Future Work

In this research, we studied how ordering biases can affect a user's mental model and reliance formation in intelligent systems and what role explanations play with such biases. Our study presents novel findings that

highlight the importance of users' first impressions on their formed mental model of the intelligent system. The results demonstrate that when encountering system strengths earlier in the usage, users built a better mental model of the system strengths as they used the system explanations more frequently. But, positive first impressions can lead to automation bias and more errors as the user is overconfident in not only the model's strengths but also the weaknesses of the system; and they generally over-rely on the system. In contrast, when encountering system weaknesses early-on, users tend to rely more on themselves and make fewer errors; likely because they develop a mental model that is skeptical of the system strengths due to their negative first impressions.

In this study, we focused on a machine learning technique that produces high-level explanations with a novice-friendly explanation interface (e.g., instead of using probabilities, we showed visual bars). While we believe our results can generalize for various real-world systems incorporating this class of explanations, these results might not generalize for low-level, more technical explanations. Future research needs to test and compare ordering bias with these explanations as well. Further, since our system employed instance-level and local explanations, additional research is needed to assess whether these results hold for higher-level, global intelligent systems.

Due to the nature of the design for our query-building tool, when users searched for an activity, we divided the video into two categories of *matched* and *not matched* based on whether the system detected the activity within each video. The detection is of course not always correct, i.e., a system might categorize a video as a *match* when the activity did *not* take place in the video (false positive error) or categorize a video as a *mismatch* while the activity is in fact taking place in the video (false negative error). For most of the activities, the number of *matched* videos was smaller than the number of *not matched* videos, and thus, users needed to explore and view fewer videos to detect false positives. Since it was easier to determine false positives, we expect that the participants would fail to catch lots of false negative errors, i.e., the videos that the system failed to match for the query. As a result, some system weaknesses were harder to identify, potentially leading to improper mental models of system weaknesses. We suspect that this is the reason the study could not find evidence of differences between the conditions based on a user's mental model of the model's weaknesses. Future research may benefit from refined evaluations focusing on both error types to test user's mental model formation for both strengths and weaknesses.

6 CONCEPTUAL MODEL OF USERS' PAST EXPERIENCES

The goal in the presented study was to investigate how anchoring bias can impact mental model formations with XAI systems. While anchoring bias is an important bias to explore, different types of experiences can introduce anchoring effects. Even still, anchoring bias is only one of the many possible types of biases. Whether conscious or unconscious, human biases initiate from the complicated human brain, where memories, past experiences, social pressures, and heuristics exist [60]. Users' backgrounds and past experiences affect how they see and experience the world and form and are affected by cognitive and societal biases. The findings of this paper, in lieu of other prior work, motivated us to dive deeper to investigate how people's collaborative efforts with intelligent tools can be affected by their past experiences. With people coming from different backgrounds, types of expertise, and social circles, it is crucial to understand how such differences can affect usage behaviours and user perceptions of the systems. To better understand possible implications for usage effects and design choices, we formalized and developed a conceptual model of users' past experiences along with associated *factors* and *outcomes* that might influence user behaviours.

In this section, we present a candidate conceptual model and explain how different categories apply to user interaction processes. We reference existing literature as examples of points for each of the categories. We will first describe the model as a whole, and will further go deeper in the details in each category.

Fig. 7. An overview of the conceptual model of users' past experiences in hte context of human-AI collaborations. The model works based on time, where each box represents time of experience: long-term past experiences, short-term past experiences, and present usage. Each of the boxes includes a set of *factors* that influence user actions that lead to certain *outcomes*. The dark purple arrows (from left to right) demonstrate how the past experiences affect those taking place more recently. As the time goes by (i.e., present turning to past), the state of each of the factors update based on more recent interactions. This constant back and forth between present, short- and long-term pasts is denoted by light purple arrows (from right to left).

6.1 Model

Here, we present the conceptual model of users' past experiences. A summarized overview of the model is seen in Figure 7. In this model, we explore how user's past experiences affect their behaviours with an intelligent system. As seen in Figure 7, we categorize the concepts based on the timeline of usage. At a high-level, we divide the time to *long-term past experiences* and *current usage*. The long-term past experiences include user backgrounds and experiences in the past, prior to using the system, while the current usage stage indicates the period since they started interacting with the tool. As people continue using the system, their recent perceptions and experiences with the tool influences their upcoming interactions and usage. Thus, current usage consists of *short-term past* and *present* phases.

For each of these categories, we include two types of elements: 1) *factors* and 2) *outcomes*. The *factors* refer to the those human behaviours that influence user choices, interactions, and decisions, ultimately leading to certain behavioral *outcomes*. Despite being referred to as outcomes, they are not easily observed nor can they easily be measured; in fact, their existence can easily be overlooked by system engineers and designers due to their abstract nature, provoking unpredictable usage behaviours. Specifically in our model, with current usage, *outcomes* from present interactions can become or contribute to the *factors* in short-term past. Changes in the *factors* in short-term past may alter the corresponding *outcomes*, and once again, the *outcomes* from short-term past are become the *factors* in the present interactions. As such, this "iterative" cycle of usage continues (flowing with the nature of time, with future turning to now, and present becoming short-term past and later, long-term past, and the shadows of the past affecting the present actions and behaviours), shaping the entirety of human-AI collaborations. This iterative process is denoted in the model with the purple arrows inside the *current usage* box. Similarly, individual past experiences and backgrounds can affect how each user perceives and utilizes AI applications. Some of these experiences and exposures to AI technology changes, the more they use such applications. These so-called "current usages" become part of the longer-term past experiences and exposures over time, with the past constantly changing and shaping the person's usage behaviours. This iterative back-and-forth is also denoted by arrows between the dotted categories (i.e., long-term past and current usage).

We continue this section by diving deeper into the main concepts of our conceptual model, referencing to a non-exhaustive list of related work in the human-centered AI and XAI research field. The summarized overview of these papers with respect to these concepts can be found in Table 1.

6.1.1 ***Long-Term Past Experiences.*** This category includes prior experiences, events, and circumstances that influence how users perceive and interact with AI systems (As summarized in Figure 7). The *factors* in this category not only causes certain behaviours during current usage, but also leads to formation of cognitive and social biases (consciously or unconsciously), not easily altered. Here, we discuss these background *factors* based on whether they were derived from personal or communal roots.

Personal *factors* originate from personal differences. A key distinctions is rooted in users' level of knowledge. While there are many ways to factor in the differences based on knowledge (such as level of education), in the XAI community, users are most commonly compared based on their level of domain expertise and familiarly with AI/ML, and in fact, many prior work has focused on studying usage behaviours based on such differences in prior

expertise. For example, in an earlier work [44], we looked into how domain experts and novices show different trends in trust calibration and their perceptions of an explainable algorithm. Our results demonstrate that domain expertise is directly associated with impression formation, which could be positive or negative depending on when users face model errors during interactions. Szymanski et al. [70] look into how different levels of users' expertise cause them to form different understandings of explanations. They found that users show different preference patterns for textual and visual explanations depending on their level of expertise. Similarly in another recent example, Ehsan et al. [16] explored how user background knowledge of AI influences their perceptions of different types of explanations, and they also found people show different behaviours towards explanation type based on their expertise in AI. All these mentioned studies, alongside many more not mentioned here, provide supporting evidence for and highlight the importance of keeping expertise and background knowledge into account when designing intelligent systems.

In the current literature, comparisons between levels of background knowledge is done in a binary fashion, where people were measured based on whether they have certain expertise or not; i.e., novices vs. the so-called experts. However, it is important to study and explore background knowledge as an spectrum and in relation with one another. For instance, the interplay between domain expertise and AI expertise could ultimately lead to certain usage behaviour. Another open question is quantifying (or measuring) expertise. A majority of papers that examine expertise as a factor rely on people's self-reported perceived expertise [16, 46] or the researchers' qualitative assessment of whether someone is an expert on a case-by-case basis [46, 70]. Such approaches propose many limitations to the scientific findings. For example, due to the Dunning-Kruger effect [13], experts might underestimate their expertise while novices may be too confident with their knowledge. Broadly speaking, people's assessment of their expertise is highly subjective and people's definitions differ when they are asked to assign ordinal numbers to their level of past knowledge. It is, thus, crucial to find more objective, standardized techniques to measure and report expertise, which presents an opportunity for future potential work.

The other personal factor from the past that we will discuss here is the memory of the past experiences and usage of intelligent systems. Aside from a mental model specifically formed for a certain AI system and how it works, people may form a broader mental model of AI technology. This mental model can be affected by personal usage experiences in the past, and whether the past encounters of AI systems lead to positive impressions of AI technology and world as a whole. However, such generalized mental models are also prune to be influenced by communal events, experiences, and circumstances. This means past experiences of others in the community can shape how people within the community perceive AI systems. For instance, hearing about negative outcomes of AI from others in social media may negatively affect people's perceptions of AI. Such societal factors may come from demographic social experiences—i.e., people from different demographic backgrounds might discern AI differently—while they may also be due to social biases (such as confirmation bias). An interesting example in this category is seen in the work by Ehsan et al [15]. They introduce a task scenario that provides examples of past decisions from prior teammates in addition to the AI predictions and explanations to facilitate users' decision-making. They describe this concept through a framework of *Social Transparency*, and discuss how presenting such historical decision-making information can shepherd positive reinforcements such as improved trust calibration and decision-making.

Overall, there still exist more open questions and future work opportunities with the communal than personal past experiences. Firstly, we still lack proper understanding of how explanations can be designed for and to benefit from communal/shared past experiences in ways that are beneficial to the users. More over, it is yet to be studied how societal and mutual understandings of intelligent systems influences current usage. While there have been prior efforts in studying decision-making provenance and historical usage summary of the same system, the implications of how others' usage and perceptions of AI technology may affect one's usage and perception of AI applications is yet to be studied. This is still an open, yet important question that needs to be answered; after all, humans are social beings and their collective past experiences and perceptions can affect them both

individually and as a group. These are complicated scenarios that cannot be conveniently controlled through a few user studies; but our incremental understanding of communal and social impacts of AI through smaller problems can help with answering broader questions.

As seen in Figure 7, while long-term past plays a significant role on how users perceive XAI systems, current usage, including the buffered past usage (or the short-term past that includes the past interactions with the tool during the current usage session) and the present interactions, determines one's perceptions and behaviours with the system. Specifically, regardless of a person's background and past experiences, an intelligent system has a chance to prove itself and its worth to the end-user during each usage session. It is during this time that the user decides whether to trust the model or not. It is then that they shape their mental models based on the predictions (recommendations) and explanations. How the model actually behaves and how the explanations are designed and included can consistently change the user's attitude towards the model. In the upcoming subsections, we will discuss the implications of current usage through an examination of short-term past and present.

6.1.2 ***Current Usage.*** When a user initially interacts with an XAI system, they only bring along those attributes associated with long-term past, as well as its outcomes (i.e., their biases). Some of the biases are linked with early-on usage of intelligent systems and can potentially determine the outcomes of the human-AI collaborations. For instance, as we observed from the results of this paper [45] and our past work [44], *first impressions* of the XAI system affects mental model formations, confidence, reliance, and trust calibration.

Another example of early-on bias is *availability heuristic*, which refers to the user's tendency to recall the information received early-on more predominantly than the information presented later-on. While such bias does not necessarily affect the user in the initial usage stage, it can be a more prominent factor as the user continues exploring and interacting with the system. In our past work [26], we found that people tend to gradually lose trust in the outcomes of a model when they were asked to provide correction feedback to model errors as early as when they started using the model. While this might seem counter-intuitive, one way of justifying it is by cognitive biases of this nature. First and foremost, we believe when people are asked to provide feedback to the model, in their mind, fixing errors might be more salient than simply observing and agreeing with the model outputs. This is an apparent example of availability bias. Moreover, as seen in our work [45], first impressions of an intelligent system can anchor users' mental model formations. So, when they are negatively anchored towards the model (i.e., by seeing model errors and having to correct them early-on), they form flawed mental models of the system. Such examples present significant issues as human-in-the-loop systems heavily rely on people's providing accurate feedback. While end users desire the ability to control the behavior of systems they rely on [68] and such feedback can successfully be used to incrementally update the model over time [17, 81], if it results in negatively biased mental models of the systems, then they may not be able to provide optimal feedback. Therefore it is vital for the AI system designers to be aware of the potential cognitive biases that may occur from including humans in the loop and consider ways to prevent their users from maintaining their initial impressions of the system, even after it has improved beyond its initial limitations.

When exploring the interplay between explainability, feedback, and human behaviours when using human-in-the-loop systems, other interesting challenges may arise. For instance, Smith-Renner et al. [67] studied whether model transparency can improve users' understanding of the model and their tendency to fix these problems, and whether asking people to provide feedback enhances their perceptions of the model and its improvements. While their user studied presented thought-provoking findings, here, we will focus on one of them. They observed that users expected the models to improve over time, even when they were not asked to provide feedback. Upon further investigations, they found that users tend to believe that ML models "get better as they function". Though their study did not investigate why this happens, they believe such misconceptions can be a result of prior experiences or general misunderstandings, supporting our argument that long-term past (either personal or communal) affect people's perception of intelligent systems.

Present usage results in numerous outcomes that are either measurable/observable behaviours (such as task accuracy, speed, interaction behaviours, and level of agreement with the model) or abstract, perceptual outcomes like trust and mental models. It is important to note that the abstract outcomes are those that are direct results of the current interaction; i.e., the most recent fragment of information that can contribute to forming mental models as opposed the complete mental model. Outcomes like trust and mental models require iterative reasoning and interactivity. They can improve or deteriorate with time and usage, and based on the other outcomes. That is why we describe an in-between stage, i.e., short-term past, that reflects the most completed version of these abstract outcomes, which we will discuss shortly.

As mental models are formed and users continue interacting with the system, the best, most current version of the user's mental model becomes a new factor that was not initially present upon the initial usage period. How users perceive the model functionality and reliability, forms expectations of the potential outcomes, and understand model weaknesses and strengths become important in the outcomes of the human-AI collaborations. Mental models can determine when users lean toward trusting the model and when not; when they agree with the system and when not; and how many errors they make in their collaborative task. A user's most recent mental model can also affect how their mental model is shaped and changed with current and future usages. While mental models have been studied intensively in other fields, such as psychology and HCI, there are still many potential opportunities to study them in the lieu of AI/XAI applications. For instance, how do users calibrate their mental models over time? What factors can influence how mental models are calibrated? Are there certain biases that can lead mental models to be calibrated towards certain perceptions? When might explanations reinforce existing biases and when might they reduce misunderstandings? There are so many open questions on how mental models are shaped and can shape people's perceptions of intelligent systems, that warrant future work in this field. We hope to have inspired future work to focus on addressing some of these challenges.

6.1.3 ***Short-Term Past Experiences.*** As users continue their session with an intelligent system, they start building a rapport with the tool. With more interactivity comes a better ability to understand and calibrate their trust and reliance through continued experiences. A user's mental models start shaping as they form their expectations and impressions of the system. Some of these impressions might be accurate while others may be flawed. It can take iterations of user interaction and machine feedback for users to solidify their mental model of any given system. The user needs opportunities to observe errors and mistakes from the system, make their own mistakes and see how the model reacts, and see what the model is capable of. The gradual development of the human-AI relationship can lead to beneficial or harmful outcomes. Much of the human-AI relationship may be formed subconsciously and reflect in the current usage behaviours. Here, we focus on two of these important outcomes: trust and reliance.

Based on observations of AI performance, users decide when to trust the model and when not to. When considered over time, users calibrate their trust to adapt to the system's most recent behavior. However, several factors may lead to mistrust in this process. Users may trust some outcomes that are incorrect or not trustworthy, or they may distrust correct output. This difficulty in effectively calibrating trust may be consequential, compounding with the effects of *long-term* past experiences to the quality and status of the formed mental models. A critical consideration here is the presence of cognitive and societal biases. In our proposed conceptual model, we highlight the importance of awareness for both *factors* and *outcomes* of past experiences related to the formation of biases and perspectives of technology. Naturally, human biases can lead to certain assumptions and conclusions in the background of their mind. When forming mental models, biases play a significant part in how perceptions, expectations, and understandings are being formed. A clear example is the findings of this paper (i.e., Nourani et al. [45]), which demonstrates how anchoring heuristic can strongly influence mental model formation, the positiveness of which relates to that of the formed first impression. From a different angle, users can form biases as a result of their interactions with the model over time. Depending on the application and the task, these biases

can potentially be of many natures, such as biases towards the future outcomes of the model (e.g., inattentional blindness, availability, hindsight, and anchoring bias), perceptions of AI and ML algorithms and technology (e.g., automation bias and expectation bias), or other people's demographic representations within the tool (societal biases). Such biases can ultimately affect current usage or even reside permanently in one's subconscious.

With improved understanding of the model intentions in different scenarios and knowing when to trust the machine outcomes, users can then decide when to rely on the model outcomes. Based on their trust, users not only decide whether to rely on the model or not, but also *when* to do so. Knowing this is not necessarily an instinct that comes easily. Rather, users need to see the model outcomes and evaluate other values that may not be easily measured directly. For instance, in scenarios with different stakeholders involved, the satisfaction of each of these stakeholders can motivate the system user to rely more on the outcomes. The user's trust, mental model of the system, and the level of reliance on the system are intrinsically intertwined.

Even though most of our discussion so far was based on first-time usage and interaction with the system in a single session, the points we made still apply to when users return to the system in a new session after having used it before. In such scenarios, the prior interactions with the model are stored in the users' mind; granted, it even becomes the case that these experiences contribute to long-term past experiences. However, when the users returns to start a new session, they have already formed perceptions of the system during the past usage. In our conceptual model, the prior experiences and perceptions of the system are shown to current present usage from the *Long-Term Past* category and *Recent Mental Model* as a factor in the *Present* category (i.e., in Figure 7). Nonetheless, we should emphasize that the user will likely have a less clear vision about the system in mind. This is the nature of the human brain: memories start fading with time. As such, upon starting a new session, the user will pick back their mental model from the past exposure to the system, but their perceptions might differ slightly from then due to the passage of time, which has similar implications for other usage factors, like trust and reliance.This passage of time has similar implications for other usage factors, like trust and reliance. Studying usage over time and over different sessions can be quite challenging; however, it is critical to understand how human usage factors calibrate over the course of time and with multiple usage sessions. In a recent attempt to study how first impressions affect users in the beginning of every usage session in a long period of time, Tolmeijer et al. [73] conduct a set of user studies that spanned across 6 days and consisting of three sessions. Their main goal was to understand how first impressions affect trust calibration over multiple sessions. This study was complementary to our prior work [44], where we found that within one session, positive first impressions of the AI model leads to higher trust and users' ability to calibrate their trust accordingly, while negative first impressions lead people to lose their trust and ability to adapt it over time. Tolmejer et al. change the accuracy of the model in each session so that each session represented either accurate or inaccurate predictions. Their results indicate that first impressions matter not just internally, but also externally across sessions, specifically when the system was often wrong (i.e., only one accurate session). Furthermore, this work demonstrates another example of how understanding model problems (i.e., user mental models of weaknesses) can impact users' trust calibration. However, given that their study was held over a longer period of time, it can further confirm the generalizability of our findings for more systems that are used over in real-world. One major difference with our prior work [44] and the work by Tolmeijer et al. [73] is that the former utilized high-fidelity explanations while the latter did not. Hence, it is yet to be explored how these findings hold with longer term usage when explanations are present, and how do explanations impact users and their first impressions when they are used over time. This is only one example open question in this area, while the study of changes and usage over time still remains an neglected area of focus, presenting opportunities for future work.

Table 1. Summary of topics from the papers discussed and presented as examples in Section 6. The topics in each column are inspired by the findings of this paper and our candid conceptual model of user's experiential biases. We carefully examined each paper to check whether each topics were heavily discussed and/or directly measured in a given paper. For the biases, we first examined whether any specific bias terms were discussed in the paper or not. In the latter case, we analyzed if the paper pays tribute to a certain bias that is not explicitly mentioned. For instance, Smith-Renner et al. [67] describe user's expectations of the AI model, which resonates with confirmation bias. While this table showcases a non-exhaustive list of recent papers to support our arguments regarding our conceptual model, by referring to this list, we can identify some open challenges that can be incremental to the findings of these papers. For instance, how does explainability help with user trust with human-in-the-loop systems? Is it capable of mitigating users' formed availability biases?

Paper	Mental Models	Trust	Explainability	Expertise	Cognitive / Societal Biases
Nourani et al. [44]	✗	✔	✔	✔	*Anchoring Bias*; *Automation Bias;*
Ehsan et al. [16]	✗	✔	✔	✔	*Automation Bias*; *Overconfidence;*
Szymanski et al. [70]	✗	✗	✔	✔	*Confirmation Bias;*
Ehsan et al. [15]	✗	✔	✔	✔	*Societal Biases;*
Tolmeijer et al. [73]	✗	✔	✗	✗	*Anchoring Bias; Reliance;*
Nourani et al. [45]	✔	✗	✔	✗	*Anchoring*; *Overconfidence; Automation Bias;*
Smith-Renner et al. [67]	✗	✔	✔	✗	*Confirmation Bias*;
Honeycutt et al. [26]	✗	✔	✗	✗	*Availability Bias;*

6.2 Takeaways and Future Work

We presented and discussed our conceptual model of users' past experiences in the context of human-AI collaborations that is mainly relying on user's prior / present experiences and impressions. We hope this organization sheds light to the current gaps and open challenges in state-of-the-art research in human-centered AI and highlights the importance of paying attention to user's prior backgrounds, experiences, and expectations. Studying human behaviours based on the timeline of usage and experiences can aid AI application designers to build systems and tools more responsibly and better thought-through. Such systems should take advantage of user differences and utilize approaches to adapt the AI technology and its outcomes to user needs and variety of behaviours. We use references from recent papers in human-centered AI community as example scientific attempts to study the concepts we present in our framework. These papers are summarized by main topics in Table 1. While we hope our candid model can be beneficial at its current state, our future plan is to continue improving it over time by performing an extensive analysis of the literature to refine the categories and organization of the topics.

7 CONCLUSIONS

In this work, we empirically explore how human subjects calibrate their trust in a policy review task with support from an XAI tool. People who saw positive system behavior early on exhibited behaviors of blind overreliance on the system output. That is, when one believes the system to perform well, they are likely to believe that the system performs well in all situations and miss the frequent mistakes. Yet, when explanations are available, this cavalier attitude is tempered and more accurate mental models develop. On the other hand, for people who saw the system perform poorly early on, the addition of explanations did little to counteract their sour impression. This is a testament to the relative weight users place on their initial experiential trust as opposed to the evidence afforded over later trials. First impressions are important in XAI systems because human biases are prevalent. This early dismissal can lead to the neglect of decision support aids and ultimately a fairly fuzzy mental model

of how the system performs. In this work, we show how poor object detection and system failures early on directly lead to underreliance and skepticism in future queries. While there was no significant difference in task completion time, people who saw the system perform poorly at first were more careful because they found the system's edge cases early on. Without clear evidence of the system's competencies, users abandoned the support provided and attempted to complete the task independent of the tool. We see that even the addition of explanation components did not help users re-calibrate their expectations or encourage them to trust the model even when it performed well. To the system designers, we recommend considering ways to balance the display of system strengths and weaknesses when users begin working with or training on intelligent tools. Our work reiterates the need to investigate additional ways to re-assess and adjust biased mental models to calibrate user expectations, build a more complete understanding of the tool's capabilities, and maintain the appropriate level of trust in system output.

As the culmination of our previous explorations into the human factors associated with explanatory systems, we conclude our paper by presenting an empirically-driven conceptual model of user's experiential biases, where we categorize human behaviours and impressions based on user experiences and backgrounds based on their time in usage. To support our discussions and our categorization, we sought after and present example related work from the human-centered AI research. We believe our organization can scaffold future design considerations for bias mitigation, prompt future research directions, and showcase methodologies from HCI/XAI that might measure these effects.

ACKNOWLEDGMENTS

This work was supported by the DARPA Explainable Artificial Intelligence (XAI) Program under award number N66001-17-2-4032 and by NSF award 1900767. The authors of this paper would like to thank the reviewers for their constructive feedback on the earlier manuscript from this paper.

REFERENCES

[1] Ashraf Abdul, Christian von der Weth, Mohan Kankanhalli, and Brian Y. Lim. 2020. *COGAM: Measuring and Moderating Cognitive Load in Machine Learning Model Explanations.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376615

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

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

[4] PAIR Team at Google Research. 2019. *People + AI Guidebook.* https://pair.withgoogle.com/chapter/mental-models/, last accessed on 12/15/2021.

[5] Moshe Bar, Maital Neta, and Heather Linz. 2006. Very first impressions. *Emotion* 6, 2 (2006), 269.

[6] Jonathan Baron. 2000. *Thinking and deciding.* Cambridge University Press.

[7] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[8] Isaac Cho, Ryan Wesslen, Alireza Karduni, Sashank Santhanam, Samira Shaikh, and Wenwen Dou. 2017. The anchoring effect in decision-making with visual analytics. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 116–126.

[9] Roberto Confalonieri, Tillman Weyde, Tarek R. Besold, and Fermín Moscoso del Prado Martín. 2021. Using ontologies to enhance human understandability of global post-hoc explanations of black-box models. *Artificial Intelligence* 296 (Jul 2021), 103471. https://doi.org/10.1016/j.artint.2021.103471

[10] Kennith J. W. Craik. 1943. *The Nature of Explanation.* Cambridge University Press. Google-Books-ID: EN0TrgEACAAJ.

[11] Julio Cesar Soares Dos Rieis, Fabrício Benevenuto de Souza, Pedro Olmo S Vaz de Melo, Raquel Oliveira Prates, Haewoon Kwak, and Jisun An. 2015. Breaking the news: First impressions matter on online news. In *Ninth International AAAI conference on web and social media*.

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

[13] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.
[14] David Dunning. 2012. Confidence considered: Assessing the quality of decisions and performance. *Social metacognition* (2012), 63–80.
[15] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.
[16] Upol Ehsan, Samir Passi, Q Vera Liao, Larry Chan, I Lee, Michael Muller, Mark O Riedl, et al. 2021. The who in explainable ai: How ai background shapes perceptions of ai explanations. *arXiv preprint arXiv:2107.13509* (2021).
[17] Ryan Elwell and Robi Polikar. 2011. Incremental learning of concept drift in nonstationary environments. *IEEE Transactions on Neural Networks* 22, 10 (2011), 1517–1531.
[18] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2 (2017), 2.
[19] Kasper Hald, Matthias Rehm, and Thomas B. Moeslund. 2019. Proposing Human-Robot Trust Assessment Through Tracking Physical Apprehension Signals in Close-Proximity Human-Robot Collaboration. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 1–6. https://doi.org/10.1109/RO-MAN46459.2019.8956335 ISSN: 1944-9437.
[20] Pamela Thibodeau Hardiman, Robert Dufresne, and Jose P. Mestre. 1989. The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition* 17, 5 (Sep 1989), 627–638. https://doi.org/10.3758/BF03197085
[21] Peter Hase and Mohit Bansal. 2020. Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior? (May 2020). https://arxiv.org/abs/2005.01831v1
[22] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).
[23] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* 25, 8 (2018), 2674–2693.
[24] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. 2019. SUMMIT: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics* (Aug. 2019), 1–1. https://doi.org/10.1109/TVCG.2019.2934659
[25] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *Proceedings of the 21st international conference on intelligent user interfaces*. 164–168.
[26] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 63–72.
[27] Philip N. Johnson-Laird. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences* 107, 43 (Oct 2010), 8. https://doi.org/10.1073/pnas.1012933107
[28] Zafar A. Khan and Won Sohn. 2011. Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elderly home care. *IEEE Transactions on Consumer Electronics* 57, 4 (Nov 2011), 1843–1850. https://doi.org/10.1109/TCE.2011.6131162
[29] Antino Kim, Mochen Yang, and Jingjng Zhang. 2020. When Algorithms Err: Differential Impact of Early vs. Late Errors on Users' Reliance on Algorithms. *Late Errors on Users' Reliance on Algorithms (July 2020)* (2020).
[30] Olga Kostopoulou, Miroslav Sirota, Thomas Round, Shyamalee Samaranayaka, and Brendan C Delaney. 2017. The role of physicians' first impressions in the diagnosis of possible cancers without alarm symptoms. *Medical Decision Making* 37, 1 (2017), 9–16.
[31] Tai Yu Lai, Jong Yih Kuo, Yong-Yi Fanjiang, Shang-Pin Ma, and Yi Han Liao. 2012. Robust Little Flame Detection on Real-Time Video Surveillance System. In *2012 Third International Conference on Innovations in Bio-Inspired Computing and Applications*. 139–143. https://doi.org/10.1109/IBICA.2012.41
[32] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
[33] Geoffrey K Lighthall and Cristina Vazquez-Guillamet. 2015. Understanding decision making in critical care. *Clinical medicine & research* 13, 3-4 (2015), 156–168.
[34] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.
[35] Stephanie M. Merritt. 2011. Affective Processes in Human–Automation Interactions. *Human Factors* 53, 4 (Aug 2011), 356–370. https://doi.org/10.1177/0018720811411912
[36] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors* 50, 2 (Apr 2008), 194–210. https://doi.org/10.1518/001872008X288574
[37] Robert K. Merton and Patricia L. Kendall. 1946. The Focused Interview. *Amer. J. Sociology* 51, 6 (May 1946), 541–557. https://doi.org/10.1086/219886
[38] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
[39] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[40] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *ACM Transactions on Interactive Intelligent Systems* (2018).
[41] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.
[42] Donald A. Norman. 1983. *Some Observations on Mental Models* (1 ed.). Lawrence Erlbaum Associates Inc. pp7-14, 7–14. https://ar264sweeney.files.wordpress.com/2015/11/norman_mentalmodels.pdf
[43] Mahsan Nourani, Donald R Honeycutt, Jeremy E Block, Chiradeep Roy, Tahrima Rahman, Eric D Ragan, and Vibhav Gogate. 2020. Investigating the importance of first impressions and explainable ai with interactive video analysis. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
[44] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.
[45] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces*. 340–350.
[46] Mahsan Nourani, Chiradeep Roy, Tahrima Rahman, Eric D Ragan, Nicholas Ruozzi, and Vibhav Gogate. 2020. Don't Explain without Verifying Veracity: An Evaluation of Explainable AI with Video Activity Recognition. *arXiv preprint arXiv:2005.02335* (2020).
[47] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The building blocks of interpretability. *Distill* 3, 3 (2018), e10.
[48] Deepak P, Sanil V, and Joemon M. Jose. 2021. On Fairness and Interpretability. *arXiv:2106.13271 [cs]* (June 2021). http://arxiv.org/abs/2106.13271 arXiv: 2106.13271.
[49] Maike Paetzel, Giulia Perugia, and Ginevra Castellano. 2020. The Persistence of First Impressions: The Effect of Repeated Interactions on the Perception of a Social Robot. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. Association for Computing Machinery, New York, NY, USA, 73–82. https://doi.org/10.1145/3319502.3374786
[50] Bjorn Petrak, Katharina Weitz, Ilhan Aslan, and Elisabeth Andre. 2019. Let Me Show You Your New Home: Studying the Effect of Proxemic-awareness of Robots on Users' First Impressions. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, New Delhi, India, 1–7. https://doi.org/10.1109/RO-MAN46459.2019.8956463 ISSN: 1944-9437.
[51] Gregory Plumb, Denali Molitor, and Ameet Talwalkar. 2018. Supervised Local Modeling for Interpretability. *arXiv:1807.02910 [cs, LG]* (July 2018). http://arxiv.org/abs/1807.02910v1 arXiv: 1807.02910 version: 1.
[52] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810 (to appear in the Proceedings of ACM CHI 2021)* (2018).
[53] Matthew Rabin and Joel L Schrag. 1999. First impressions matter: A model of confirmatory bias. *The quarterly journal of economics* 114, 1 (1999), 37–82.
[54] Tahrima Rahman, Prasanna Kothalkar, and Vibhav Gogate. 2014. Cutset networks: A simple, tractable, and scalable approach for improving the accuracy of Chow-Liu trees. In *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 630–645.
[55] William E Remus and Jeffrey E Kottemann. 1986. Toward intelligent decision support systems: An artificially intelligent statistician. *MIS Quarterly* (1986), 403–418.
[56] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
[57] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High Precision Model-Agnostic Explanations. *Association for the Advancement of Artificial Intelligence (www.aaai.org)* (2018), 9.
[58] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*. Springer, 184–195.
[59] Chiradeep Roy, Mahesh Shanbhag, Mahsan Nourani, Tahrima Rahman, Samia Kabir, Vibhav Gogate, Nicholas Ruozzi, and Eric D Ragan. 2019. Explainable Activity Recognition in Videos.. In *IUI Workshops*.
[60] Charlotte Ruhl. 2021. *Cognitive Bias Examples*. www.simplypsychology.org/cognitive-bias.html, last accessed on 12/23/2021.
[61] J. Edward Russo, Eric J. Johnson, and Debra L. Stephens. 1989. The validity of verbal protocols. *Memory & Cognition* 17, 6 (Nov 1989), 759–769. https://doi.org/10.3758/BF03202637
[62] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I can do better than your AI: Expertise and explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 240–251.
[63] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 618–626. https://doi.org/10.1109/ICCV.2017.74 ISSN: 2380-7504.
[64] Rachel Benish Shirley and Carol Smidts. 2018. Bridging the simulator gap: Measuring motivational bias in digital nuclear power plant environments. *Reliability Engineering & System Safety* 177 (Sept. 2018), 191–209. https://doi.org/10.1016/j.ress.2018.04.016

[65] Ben Shneiderman. 2020. Bridging the Gap Between Ethics and Practice: Guidelines for Reliable, Safe, and Trustworthy Human-centered AI Systems. *ACM Transactions on Interactive Intelligent Systems* 10, 4 (Dec. 2020), 1–31. https://doi.org/10.1145/3419764

[66] Winston R. Sieck and Hal R. Arkes. 2005. The recalcitrance of overconfidence and its contribution to decision aid neglect. *Journal of Behavioral Decision Making* 18, 1 (Jan. 2005), 29–53. https://doi.org/10.1002/bdm.486

[67] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S Weld, and Leah Findlater. 2020. No explainability without accountability: An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[68] Simone Stumpf, Erin Sullivan, Erin Fitzhenry, Ian Oberst, Weng-Keen Wong, and Margaret Burnett. 2008. Integrating rich user feedback into intelligent user interfaces. In *Proceedings of the 13th international conference on Intelligent user interfaces*. 50–59.

[69] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[70] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[71] Dairazalia Sánchez, Monica Tentori, and Favela Jesús. 2008. Activity Recognition for the Smart Hospital. *IEEE Intelligent Systems* 23, 02 (Apr 2008), 50–57. https://doi.org/10.1109/MIS.2008.18

[72] Philip E Tetlock. 1983. Accountability and the perseverance of first impressions. *Social Psychology Quarterly* (1983), 285–292.

[73] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second Chance for a First Impression? Trust Development in Intelligent System Interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87.

[74] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. 2018. Suspicious human activity recognition: a review. *Artificial Intelligence Review* 50, 2 (Aug 2018), 283–339. https://doi.org/10.1007/s10462-017-9545-7

[75] Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, and Ioana Cocu. 2021. Moral Decision Making in Human-Agent Teams: Human Control and the Role of Explanations. *Frontiers in Robotics and AI* 8 (May 2021), 640647. https://doi.org/10.3389/frobt.2021.640647

[76] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A human-centered agenda for intelligible machine learning. *Machines We Trust: Getting Along with Artificial Intelligence* (2020).

[77] Emily Wall, Leslie Blaha, Celeste Paul, and Alex Endert. 2019. A formative study of interactive bias metrics in visual analytics using anchoring bias. In *IFIP Conference on Human-Computer Interaction*. Springer, 555–575.

[78] David S Watson, Jenny Krutzinna, Ian N Bruce, Christopher EM Griffiths, Iain B McInnes, Michael R Barnes, and Luciano Floridi. 2019. Clinical applications of machine learning algorithms: beyond the black box. *Bmj* 364 (2019).

[79] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.

[80] Jin Xu and Ayanna Howard. 2018. The Impact of First Impressions on Human- Robot Trust During Problem-Solving Scenarios. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. 435–441. https://doi.org/10.1109/ROMAN.2018.8525669 ISSN: 1944-9437.

[81] Koichiro Yamauchi. 2009. Optimal incremental learning under covariate shift. *Memetic Computing* 1, 4 (2009), 271.

[82] Serena Yeung, Francesca Rinaldo, Jeffrey Jopling, Bingbin Liu, Rishab Mehra, N. Lance Downing, Michelle Guo, Gabriel M. Bianconi, Alexandre Alahi, Julia Lee, and et al. 2019. A computer vision system for deep learning-based detection of patient mobilization activities in the ICU. *npj Digital Medicine* 2, 11 (Mar 2019), 1–5. https://doi.org/10.1038/s41746-019-0087-z