

Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems

Mahsan Nourani
University of Florida
mahsannourani@ufl.edu

Donald R. Honeycutt
University of Florida
dhoneycutt@ufl.edu

Chiradeep Roy
University of Texas in Dallas
cxr161630@utdallas.edu

Tahrima Rahman
University of Texas in Dallas
tahrima.rahman@utdallas.edu

Jeremy E. Block
University of Florida
j.block@ufl.edu

Eric D. Ragan
University of Florida
eragan@ufl.edu

Vibhav Gogate
University of Texas in Dallas
vibhav.gogate@utdallas.edu

ABSTRACT

EXplainable Artificial Intelligence (XAI) approaches are used to bring transparency to machine learning and artificial intelligence models, and hence, improve the decision-making process for their end-users. While these methods aim to improve human understanding and their mental models, cognitive biases can still influence a user's mental model and decision-making in ways that system designers do not anticipate. This paper presents research on cognitive biases due to ordering effects in intelligent systems. We conducted a controlled user study to understand how the order of observing system weaknesses and strengths can affect the user's mental model, task performance, and reliance on the intelligent system, and we investigate the role of explanations in addressing this bias. Using an explainable video activity recognition tool in the 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. Our work presents strong findings that aim to make intelligent system designers aware of such biases when designing such tools.

CCS CONCEPTS

- Human-centered computing → Empirical studies in HCI; User studies;
- Computing methodologies → Neural networks.

KEYWORDS

Explainable AI, Cognitive Biases, HCI, User Studies

ACM Reference Format:

Mahsan Nourani, Chiradeep Roy, Jeremy E. Block, Donald R. Honeycutt, Tahrima Rahman, Eric D. 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 (IUI '21), April 14–17, 2021, College Station, TX, USA*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3397481.3450639>

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 *blackbox* algorithms are hard to understand for the users due to the complexity of their inner logic [47]. 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 [4].

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 [27]. Explanations can be anything from general information about and extracted from the model (e.g., post-hoc explanations [21]) to annotation of the input to highlight the features used in the decision-making process (e.g., [36]). 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 [48]. 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

IUI '21, April 14–17, 2021, College Station, TX, USA
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8017-1/21/04...\$15.00
<https://doi.org/10.1145/3397481.3450639>

by previous work in psychology (e.g., [14]). 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 [31]. 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 work 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.

2 RELATED WORK

Researchers in the human-computer interaction (HCI) community have been studying XAI systems from different angles. There have been various research on design guidelines and reviews for explanations based on their scope [1, 2], type [19], and target users [7, 40]. In the visualization community, researchers implemented and discussed different visualization techniques to improve user understanding of the model as a whole (e.g., [12, 32]), i.e., providing a global overview (or explanations) for how the model works. However, most of the work in the HCI, machine learning, and artificial intelligence communities have been focused on local explanations that explain model behavior for each input-output [1]. In this paper, we will also focus on local explanations.

The visualization aids used to describe system performance serve as the basis for users to construct their understanding—or mental model—of the system limits and competencies. Previous research shows that it is not easy to capture and measure mental models due to their temporal nature and their influence on user disposition [29].

Since their initial description in 1943 [6], mental models are generally inferred from a variety of user study techniques, such as think-aloud approaches [39], interviews [24], and well-constrained survey questionnaires [10, 11]. 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 [22, 23]. 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., [11, 27]). In AI/XAI research, prediction tasks are commonly employed after users have time to observe how the system performs [11, 40]. 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. [33] 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 [14]; 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.

Researchers in HCI and psychology communities have been studying different cognitive and heuristic biases over the years. In his book, Baron [3] 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 be a basis for decision-making. Regardless, psychologists have found that people often adapt their beliefs as they are reluctant to face any consequences for these beliefs. One of the biases under this category is *primacy effect*, which is a similar bias to what we are addressing in this paper—also studied under different names (e.g., anchoring bias [5, 45], order bias [35], and first impressions [17, 31]). With this bias, a person's initial assumptions or impressions might affect their future behaviors. In intelligent systems, these first impressions can affect future behaviors and hence, decisions [31]. In a survey about decision-making in critical systems, Lighthall and Vazquez-Guillemet [20] discuss a few of these heuristic biases that can affect a user's decisions. For instance, *confirmation bias* refers to when a person tries to collect redundant information to find more evidence that their initial assumption is correct. They argue two causes for this bias: (1) anchoring bias, where a person's decision on some variable is biased based on another variable; and (2) a psychological tendency to rely on a[n incorrect] decision they already made rather than restarting their decision-making process. This is similar to what we are exploring in this paper, in the sense that we are studying the mental models after the first impressions are formed, as we believe first impressions can anchor a person's decision and prevent them from changing it in the future. Referring to this paper, Wang et al. [46] proposed a theory-driven framework to link explanation design to user reasoning goals in order to mitigate the cognitive and some of the heuristic biases that can affect the decision-making process.

In recent work, Kim et al. [16] show and discuss that the time *when* the 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. Our work is similar, but extends this examination of first impressions by exploring the interplay of model transparency on mental models and user reliance. In other recent work, Nourani et al. [31] studied how domain expertise affects users’ first impression formations of an intelligent system, and how these impressions impact their trust and its evolution over time. Through a user study with a simulated classification scenario, they found that first impressions can significantly affect trust when users are familiar with the domain; that is, with negative first impressions, users had significantly lower trust while their naive counterparts adjusted their level of trust based on their observations of the system. Our work, however, focused on the user’s mental model in a more exploratory scenario, meaning that we have less control over what a user observes and how they experience 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 [18], airport security [44], smart hospitals [42, 49], and elderly care [15]. 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 [37]. 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 [38]. 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 [37], 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 [41] 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 [34] that modeled a conditional probability distribution of the ground (true) labels given the noisy labels from the neural network, i.e., $P(G^{1:t}|E^{1:t})$ where $G^t = \{G_1^t, \dots, G_n^t\}$ is the set of ground labels at frame t and $E^t = \{E_1^t, \dots, E_n^t\}$ 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 a video activity searching tool to allow users to build specific queries and sort the videos from the dataset. In this 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

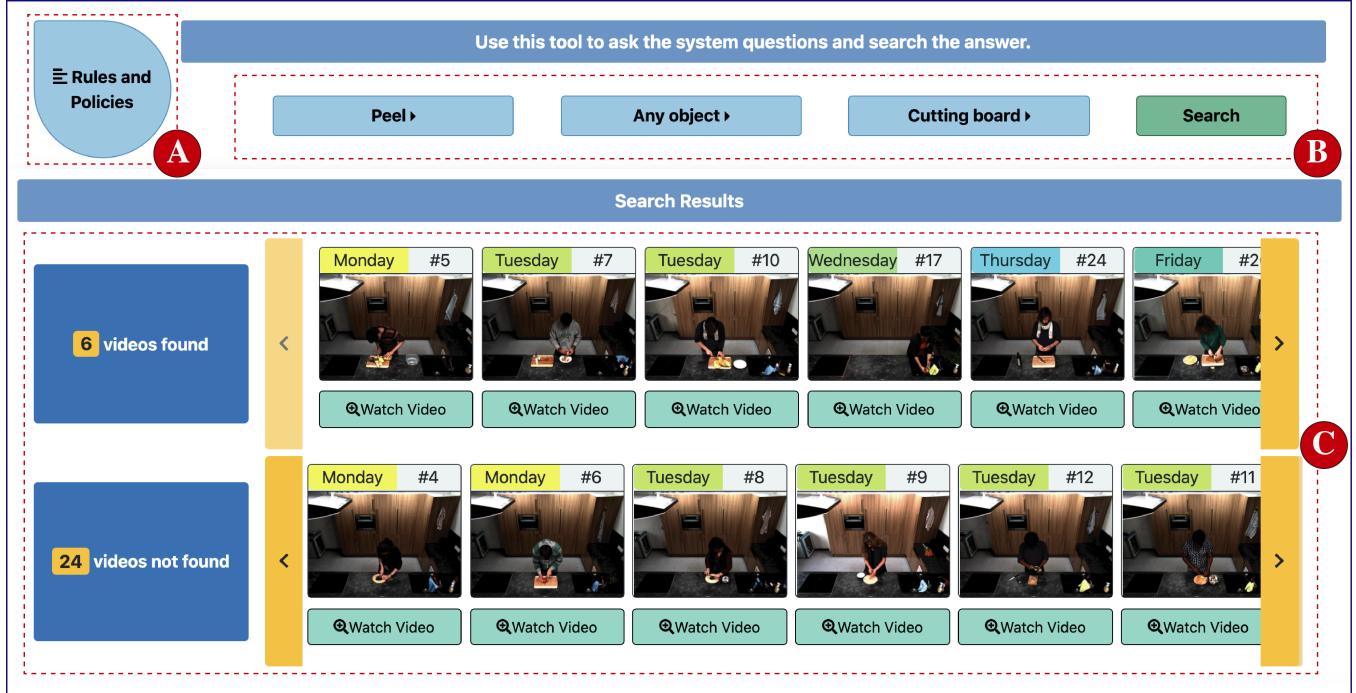


Figure 1: The main overview of the user interface. By clicking in the top left corner (A), a panel opens from the left side of the screen that includes a list of policies. Here, users recorded the kitchen's compliance with each statement. (B) Users selected components from three drop-downs to build a 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.

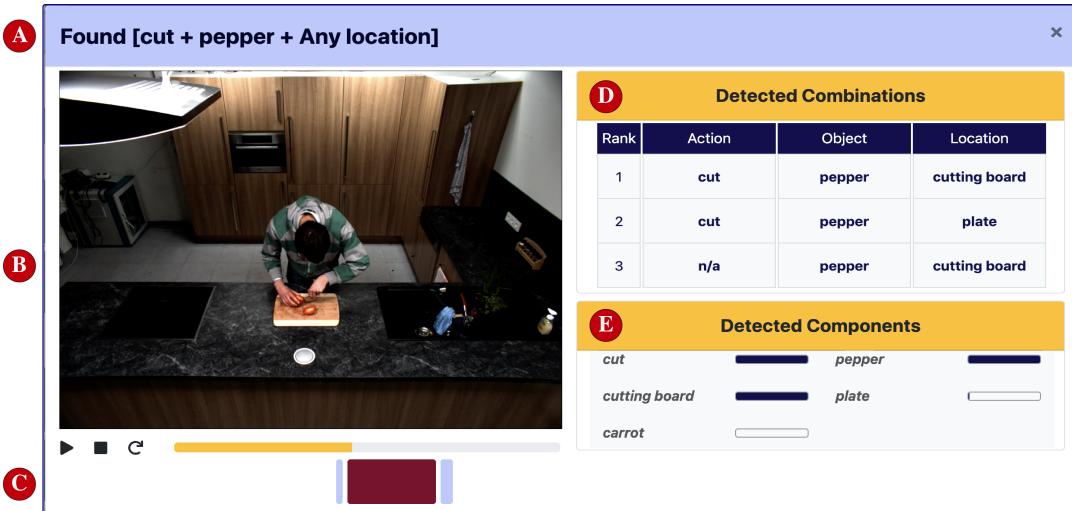


Figure 2: When clicking on the *watch video* button in the main interface, as seen in Fig. 1, participants would see a modal to allow them to watch the video. (A) showed the selected query and whether the query is found or not found in the video (B). If they were in the *explanation presence*, they were 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).

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 [9], 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 *pineapple* 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.

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., [13, 27, 31]). With a slider, users indicated

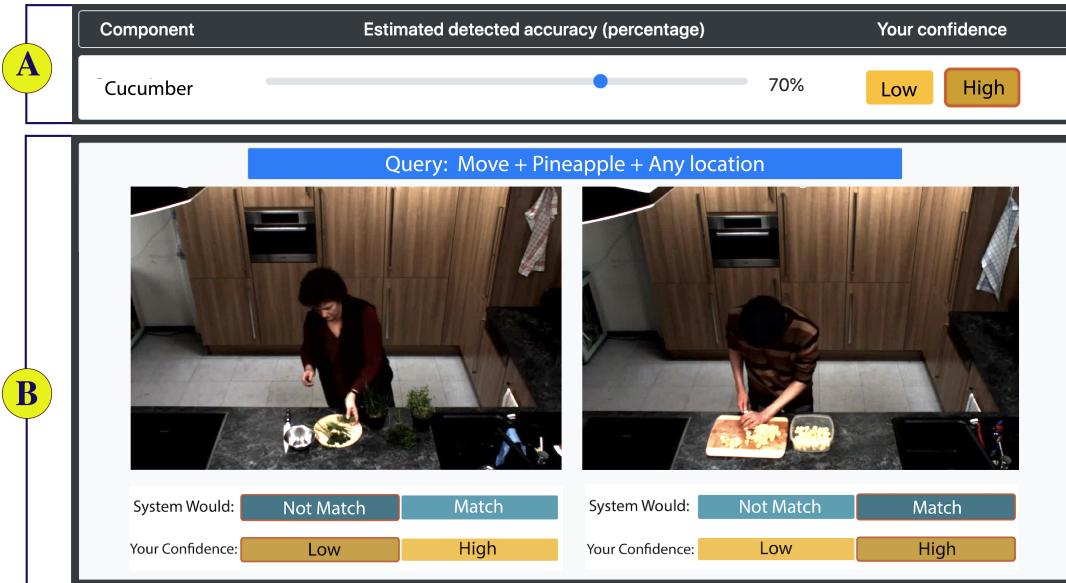


Figure 3: Examples of the mental model questions for the user study. (A) The user estimated the accuracy for cucumber was 70% and had a *high* confidence in their estimation. (B) Frame-query estimation where the user guessed whether the system matched each frame to the query and rated their confidence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

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 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 [11, 27].

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.

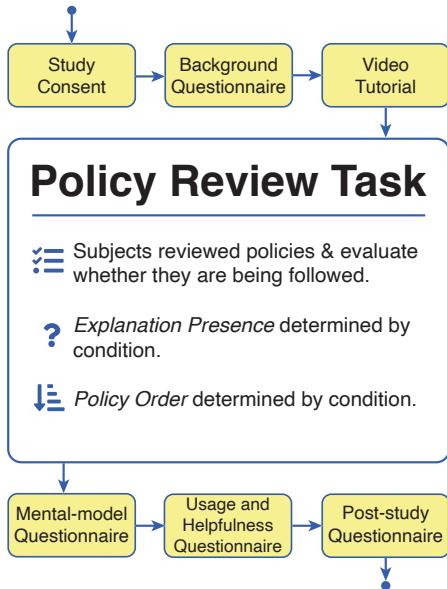


Figure 4: An overview of the study procedure.

4 RESULTS

In this section, we present the measures of our study and provide an analysis of the results. Some of our findings were previously accepted and presented in an extended abstract [30] in ACM CHI 2020. The current paper provides a more in-depth analysis of those measures, as well as measures not previously presented.

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. 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:

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

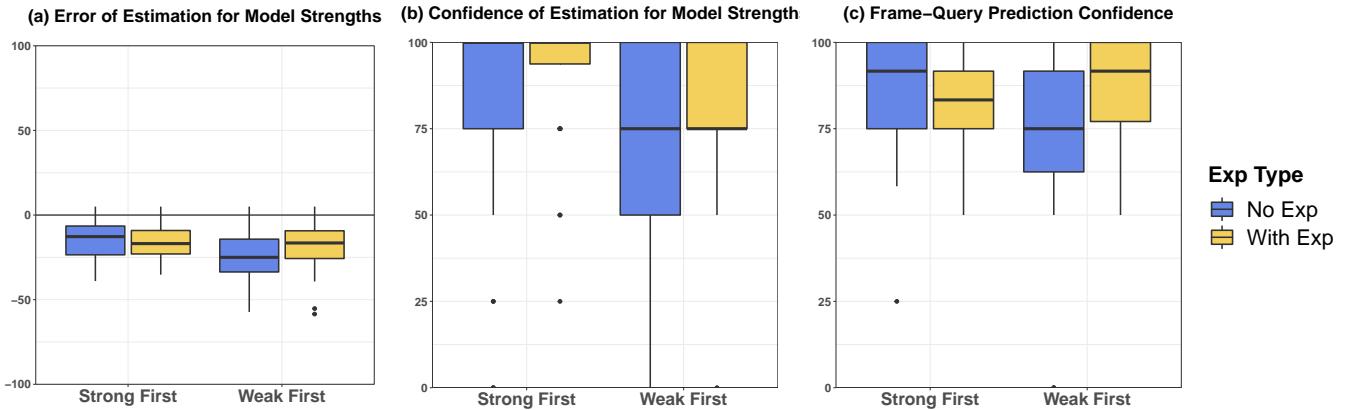


Figure 5: Mental model metrics. (a) Participants’ error of estimation for component accuracy (below 0 is underestimation). (b) Percentage of components for which participants rated as being confident in their estimation. (c) Percentage of frame-query pairs for which participants felt confident in their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

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 ($p < 0.01$). 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.

We designed the frame-query prediction task to measure the user’s granular mental model based on the specifics of the system.

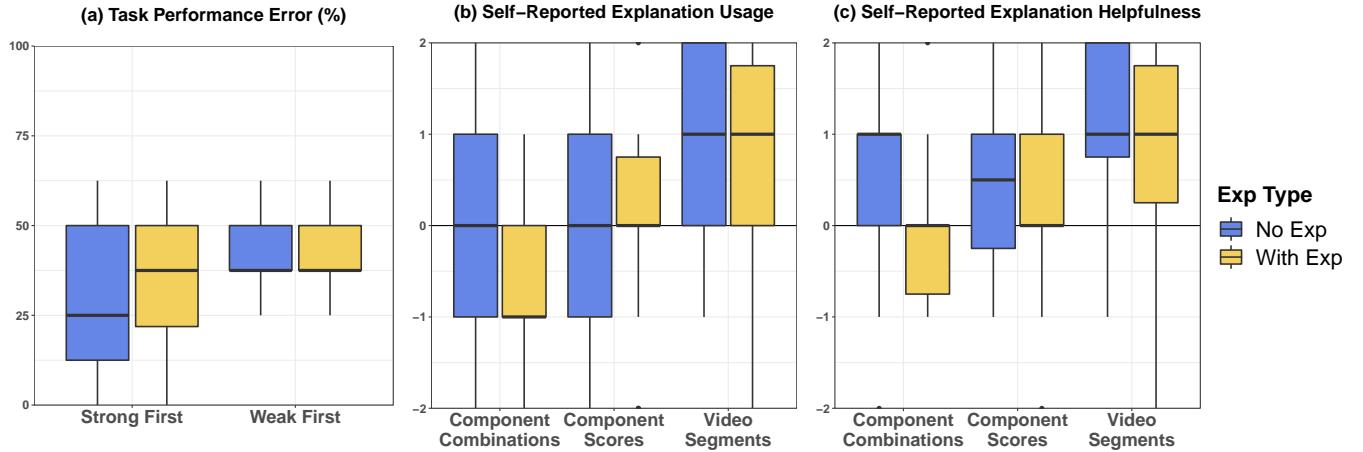


Figure 6: Reliance and Usage metrics. (a) Participant error on the policy task (Percentage). (b) Responses to the question "How much did you use this element?". (c) Responses to the question "How helpful did you find this element?". The last two were measured on a 5-point Likert scale, with higher values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

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 misbelief 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 [8, 28]. Similarly, previous research suggests overreliance can cause several problems [4, 40], 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 [43], 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., [1, 7, 25, 26]) 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 Implications for Intelligent System Designers

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 [27]. 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 [1, 7, 19]) 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 Conclusion

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.

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.

REFERENCES

- [1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
- [2] 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.
- [3] Jonathan Baron. 2000. *Thinking and deciding*. Cambridge University Press.
- [4] 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.
- [5] 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.
- [6] Kenneth J. W. Craik. 1943. *The Nature of Explanation*. Cambridge University Press. Google-Books-ID: EN0TrgEACAAJ.
- [7] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).
- [8] David Dunning. 2012. Confidence considered: Assessing the quality of decisions and performance. *Social metacognition* (2012), 63–80.
- [9] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web 2* (2017), 2.
- [10] 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>
- [11] 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).
- [12] 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.
- [13] 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.
- [14] 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>
- [15] 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>
- [16] Antino Kim, Mochen Yang, and Jingjing 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).
- [17] 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.
- [18] 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>
- [19] 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.
- [20] Geoffrey K Lighthall and Cristina Vazquez-Guillamet. 2015. Understanding decision making in critical care. *Clinical medicine & research* 13, 3-4 (2015), 156–168.
- [21] Zachary C Lipton. 2018. The mythos of model interpretability. *Queue* 16, 3 (2018), 31–57.
- [22] Stephanie M. Merritt. 2011. Affective Processes in Human-Automation Interactions. *Human Factors* 53, 4 (Aug 2011), 356–370. <https://doi.org/10.1177/0018720811411912>
- [23] 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>
- [24] 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>
- [25] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.
- [26] 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).
- [27] Sina Mohseni, Nilofar Zarei, and Eric D Ragan. 2018. A survey of evaluation methods and measures for interpretable machine learning. *ACM Transactions on Interactive Intelligent Systems* (2018).
- [28] Don A Moore and Paul J Healy. 2008. The trouble with overconfidence. *Psychological review* 115, 2 (2008), 502.
- [29] 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
- [30] 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.
- [31] 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.
- [32] 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.
- [33] Forough Pourabzi-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).
- [34] 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.
- [35] William E Remus and Jeffrey E Kottemann. 1986. Toward intelligent decision support systems: An artificially intelligent statistician. *MIS Quarterly* (1986), 403–418.
- [36] 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.
- [37] 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.
- [38] 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*.
- [39] 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>
- [40] 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.
- [41] 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.
- [42] 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>
- [43] Philip E Tetlock. 1983. Accountability and the perseverance of first impressions. *Social Psychology Quarterly* (1983), 285–292.
- [44] 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>
- [45] 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.
- [46] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.
- [47] David S Watson, Jenny Krutznick, 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).
- [48] Daniel S Weld and Gagan Bansal. 2019. The challenge of crafting intelligible intelligence. *Commun. ACM* 62, 6 (2019), 70–79.
- [49] 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>