Investigating the Importance of First Impressions and Explainable AI with Interactive Video Analysis

Mahsan Nourani
University of Florida
mahsannourani@ufl.edu

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

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

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

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

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

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

CHI '20 Extended Abstracts, April 25–30, 2020, Honolulu, HI, USA.

ACM ISBN 978-1-4503-6819-3/20/04.
http://dx.doi.org/10.1145/3334480.3382967

Abstract
We present research on how the perception of intelligent systems can be influenced by early experiences of machine performance, and how explainability potentially helps users develop an accurate understanding of system capabilities. Using a custom video analysis system with AI-assisted activity recognition, we studied whether presenting explanatory information for system outputs affects user perception of the system. In this experiment, some participants encountered AI weaknesses early, while others encountered the same limitations later in the study. The difference in ordering had a significant impact on user understanding of the system and the ability to detect AI strengths and weaknesses, and the addition of explanations was not enough to counteract the strong effects of early impressions. The results demonstrate the importance of first impressions with intelligent systems and motivate the need for improved methods of intervention to combat automation bias.

Author Keywords
Human-centered Machine Learning; Explainable Machine Learning; Empirical User Studies;

CCS Concepts
•**Human-centered computing** → **Empirical studies in HCI;** *User studies;*

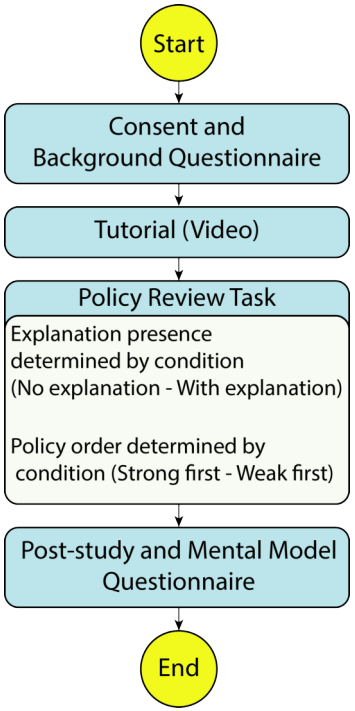

Figure 1: Summary of the study procedure. The policy review task differed between participants based on their condition.

Introduction

Intelligent systems incorporate machine learning and artificial intelligence (AI) algorithms to help their users with certain tasks and the decision-making process [3, 14, 16]. However, users of such systems often find it hard to understand how such systems work, why they are showing certain behaviors and outputs, and what they are trying to achieve. In an attempt to solve this problem, researchers propose adding explanations to these algorithms.

Explanations can be used in different contexts and have different scopes. *Local explanations* aim to explain and justify a system's rationale at the output-level, i.e., they explain why each output is generated [8, 12, 9]. *Global explanations*, on the other hand, aim to explain how a model works from a higher level, for instance, by visualizing the layers in a neural network [9, 1, 7]. As global explanations aim to represent a model as a whole, they might show both system strengths and weaknesses at the same time, with the goal of helping users build a more accurate mental model of the system. However, in practice, it is not always feasible to provide global explanations [1]. Explainable system designers, therefore, use local explanations, which allow users to build an appropriate mental model of the system by gaining experience with the system over time. As a result, the order through which users encounter system outputs can play an important role in how accurate their final mental model of the system will be.

Related to ordering, prior research has demonstrated that *primacy effects* can influence how impressions are formed [2, 5]. Studies showed that participants who receive positive information first tend to focus on more positive features when describing a context [15]. In a recent study, Rey et al. [11] found strong evidence that order through which output is retrieved in a comparison with large amounts of information influences human's decision-making process, even when the number of negative and positive features are similar. In a human-robot interaction study, Xu and Howard [20] showed that users trust a robot more when the robot provides correct advice in the beginning. This phenomenon has been studied by researchers from different communities and under different names, such as anchoring bias [17, 18, 4, 19].

In this paper, we present a user study with an open-ended task scenario involving video analysis with an AI activity recognition system. We tested how the order of the observed weaknesses and strengths can affect users mental model and task-performance with an explainable intelligent system. The results show that first impressions with a system can significantly affect user's task-error and perception of the system accuracy. In the tested context, the addition of explanations was not enough to counteract the strong effects of early impressions.

Method

In this experiment, we aimed to study how the addition of explanations and a user's first impression of an XAI (i.e., eXplainable AI) system would affect their task performance and mental model of the system. We hypothesized that the presence of explanations would increase user task performance while communicating the competencies and limitations of the AI system. We also expected that encountering more system weaknesses early on would lead to less confidence in, and less reliance on the outputs of an XAI system compared to early experiences demonstrating reliable system performance.

XAI System and User Task

The XAI system was trained to identify activities in cooking videos from the TACoS Multi-level dataset [10]. Figure 2

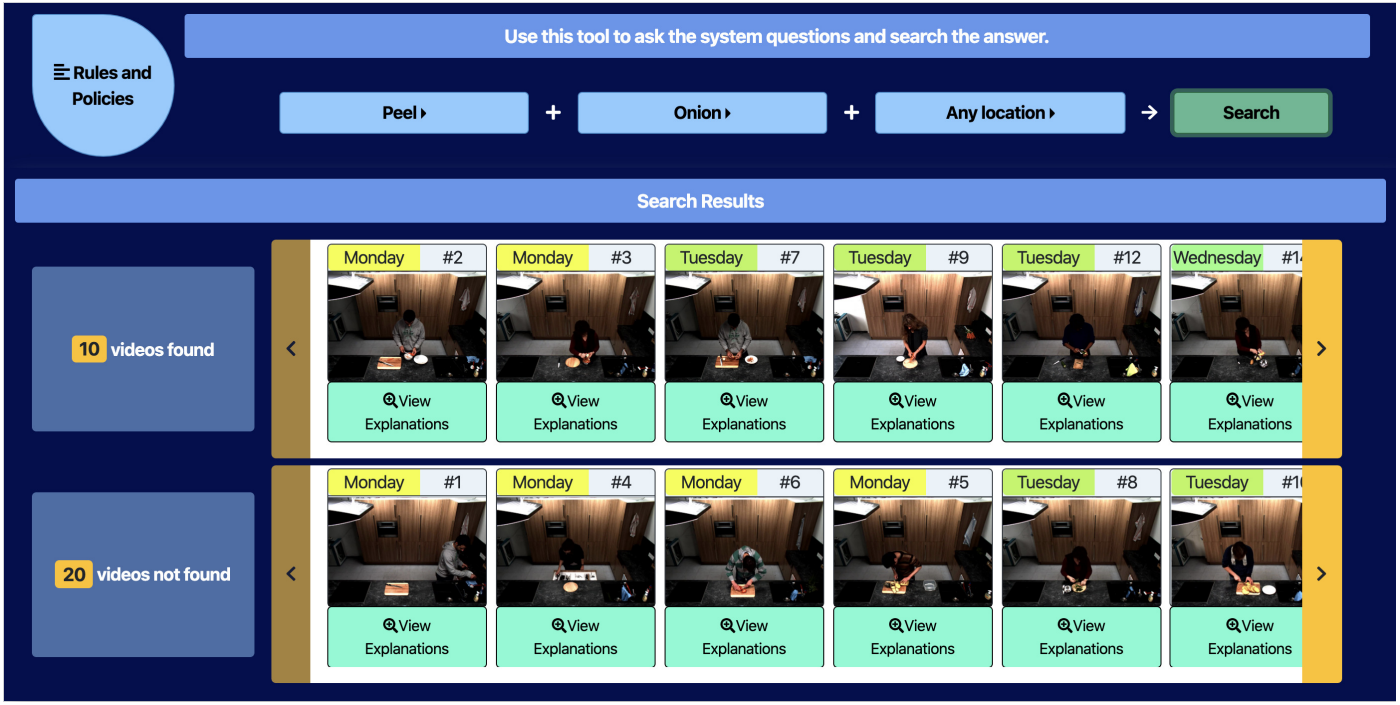

Figure 2: An overview of the interface after a user queried for videos in which "**Onion** is being **peel**ed at **any location**".

	With Exp	No Exp
Strong First	28	29
Weak First	28	29

Table 1: Number of participants in each condition.

shows the system's user interface for the study. More details about the training and model are available in an earlier workshop paper [13], while the current paper presents an updated system interface and a new user study.

To assess a user's mental model of the XAI system, participants first need to build a mental model by interacting with the system over time. Thus, we designed an experimental task where participants were given a set of 30 videos, each tagged with a day of the week (Monday to Friday). They were further given nine kitchen policies (e.g., "Employees must not use pineapples more than three days a week.") and had to determine, through video review, whether each policy was followed by the employees or not.

To assist with this task, the XAI system allowed users to query certain combinations of actions, objects, and locations, i.e., an open-ended task. When users searched for a query, the system would show a list of videos that matched the query and a separate list for all other videos. For each video, the system showed a thumbnail of the video frame that is most relevant to the query, if any of the query com-

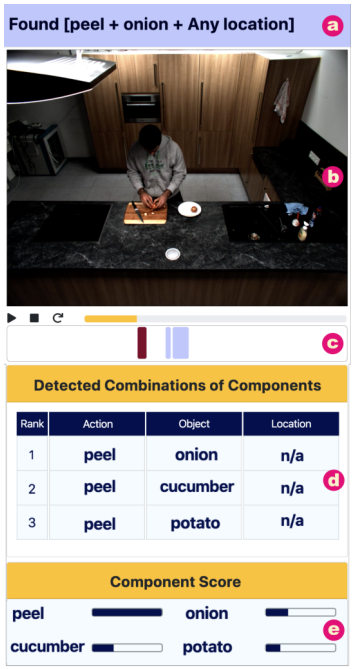

Figure 3: Explanation details shown to users upon clicking on a video from the list of Figure 2. (a) whether the video matches the searched query, (b) the video player, (c) the most relevant video segments to the query, (d) the top 3 combinations of components detected for the selected segment, and (e) the system's confidence that each component is present in the selected segment. Users without explanations only saw (a) and (b).

ponents were found. Otherwise, the thumbnail showed the middle frame of the video.

Figure 2 shows the interface of our system. The rules and policies button in the top left showed the policies a participant was asked to verify. Participants could select an action, object and location from the three drop-down lists at the top to query the system. For the depicted query *[Peel + Onion + Any location]*, the system returned 10 matching videos and 20 non-matching videos. Clicking the green button below a video opened the video player with that video for further explanations on why the clicked video was categorized as a hit or not. Figure 3 shows an example of this detail view. The system highlights time segments of the video relevant to the system's answer for the query. Upon clicking each of these segments, the system showed the top three combinations of components detected together and the individually detected components with a confidence rating.

Study Design

To test our hypotheses, the study followed a 2x2 between-subjects design with two independent variables: (1) policy order and (2) explanation presence.

Of the nine kitchen policies, the *policy order* factor determined if participants saw system strengths or weaknesses early in the study. Four of the policies asked about activities the system correctly classified (strengths) while four policies focused on activities that the system often returned superfluous incorrect positive matches or failed to match a policy's counterexample. In addition, one easy-to-confirm policy was consistently used as an attention check since participants were unsupervised for their task.

As a separate experimental factor, *explanation presence* determined if a participant saw explanations or not. While all the participants saw the same set of policies and main interface as seen in Figure 2, participants in the *no explanation* conditions only observed the video player and query information upon clicking on a video thumbnail (Figure 3a and 3b), while those *with explanation* saw the explanations as well (Figure 3c, 3d, 3e). To avoid learning effects across conditions, the 2x2 study was conducted between subjects (i.e., a total of four conditions where each participant completed one condition). Participants were randomly assigned to one of the four conditions at the beginning of the study before the main task.

Participants and Procedure

We recruited 120 university students to participate in our study, of which 114 passed the attention check. Table 1 shows the number of participants across conditions. Participants generally completed the experiment in a single, one-hour session and were asked to use the custom web application on their personal laptop or desktop computer (the interface components for the main task are shown in Figures 2 and 3).

Figure 1 shows a summary of the study procedure. First, participants completed a background questionnaire where they reported demographic information as well as reporting their level of comfort with machine learning before watching a tutorial video that described how to use the system to asses the set of policies. Users were free to answer the policies in any order, though in-lab pilot testing indicated that participants generally reviewed policies in order from top to bottom. While it was common to start at the top of the list, participants were further encouraged to follow this top-down progression by a video tutorial mimicking this behavior. After providing answers (yes or no) for all policies, participants answered a post-study questionnaire, which included questions to test participant understanding of the

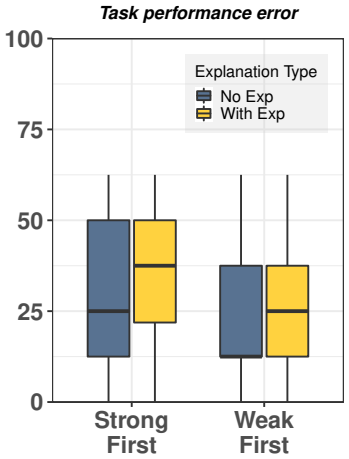

Figure 4: Participant task error by condition (Percentage).

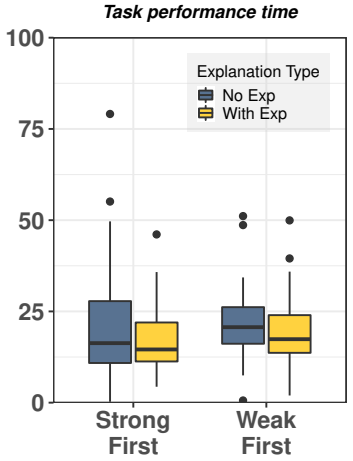

Figure 5: Participant task time by condition (Minutes).

system's ability to detect certain components (objects, actions, and locations) in the videos.

Results

In this section, we describe the measures in the study and the analysis. For each metric, we performed an independent two-way factorial ANOVA.

User-Task Performance

To address our first hypothesis regarding user-task performance, we collected the total time spent on the policy-review task and the number of falsely answered policies. For data cleaning from the online study, we removed any period of inactivity longer than five minutes, which were found through the interaction logs. While the two-way ANOVA did not show a significant effect for the presence of explanations, participants who observed system weaknesses first had significantly less error in their answers to the policy questions than participants who encountered system strengths first, with $F(1, 106) = 6.55, p < 0.05$.

No evidence of an interaction effect between explanation presence and policy order was observed. Additionally, no significant effects were observed on completion time. Figures 4 and 5 show the distribution of these results across the conditions.

Perceived Component Accuracy

After participants finished the policy-review task, they provided estimates of the AI's detection accuracy for a given set of components from the videos. We chose 9 components (8 components explicitly mentioned in the policies and 1 that was not), and participants separately estimated the detection accuracy for each as a percentage. Furthermore, they were asked to indicate their confidence in each of their estimations (low or high confidence). We selected the components so that five corresponded to system weaknesses (low AI accuracy) and four for detection strengths (high AI accuracy).

For the analysis purposes, we used the error of the average for both weaknesses and strengths for each participant separately, i.e., two metrics per participants. A similar approach was used for the confidence scores. For system weaknesses, the statistical tests did not indicate significant effects for accuracy or confidence. For the system strengths, however, participants who observed weaknesses first were shown to underestimate the system's accuracy significantly more than participants who saw strengths first, with $F(1, 106) = 6.24, p < 0.05$. Additionally, participants who observed weaknesses early on were significantly less confident about their estimations compared to those who saw strengths early, with $F(1, 106) = 3.94, p < 0.05$. No evidence was found of any effect of explanation presence on perception of accuracy. Figures 6 and 7 show the distribution of these results across the conditions.

Usage of Explanations

Finally, at the end of the study and only for the participants with explanations, we asked them to report how useful they found each of the explanation types (Figure 3c, 3d, and 3e) on 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, and then, performed a two-way independent ANOVA on explanation usage. The results show participants who encountered weaknesses first used system explanations significantly less than participants who encountered strengths first, with $F(1, 156) = 4.76, p < 0.05$.

Discussion

Our goal for this research was to explore the effects of explanation presence and order of encountering system

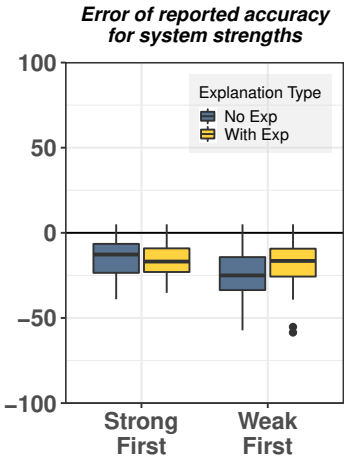

Figure 6: Error of reported accuracy for system strengths (Percentage Error).

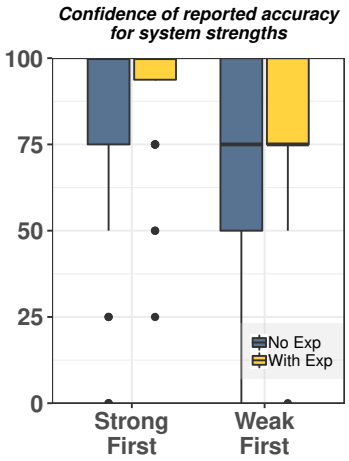

Figure 7: User confidence of reported accuracy for system strengths (Percentage).

weaknesses and strengths on user mental model and task performance in an intelligent system. According to the study results, participants who observed weaknesses first were able to complete the task with less error than those who saw strengths first. This indicates that first impressions of an intelligent system could lead to an effect of automation bias [6], a situation in which users rely on and favor the outputs of an automated system in a decision-making scenario over other contradictory information. Participants who saw strengths first were more susceptible to relying on the system's answers even when they were incorrect. This aligns with previous research on primacy effects [2, 11], in that the user's first impression of the system dictates their level of reliance on its abilities.

The primacy effect appeared to also be present in the participants' interpretations of the system accuracy. We expected participants to underestimate the accuracies of system strengths that were relatively high (each above 92%). However, participants in weak-first conditions underestimated the accuracy of system strengths significantly more compared to their counterparts. Even while observing system strengths, it appears that their initial impressions of the system led them to believe these components were weak as well.

Experiencing a positive first impression seems to lead participants to rely on the system's outputs, even at the times the system is not correct. This observation can be explained by the automation bias, which can cause larger error during the decision-making process, as backed up by our results. On the other hand, more reliance on the system also increased the usage of system explanations and allowed for a better judgement of the system strengths. From another perspective, a negative first impression prevents users from relying on the system outputs, creating an insufficient understanding of the system capabilities and reducing the usage of explanations. However, in cases of AI detection failures, these participants were not misled by the system outputs and were less prone to error caused by automation bias.

Overall, these findings highlight the importance of first impressions for users interacting with explainable AI systems, as first impressions (good or bad) can affect a user's behaviors with the system. A positive first impression might invoke automation bias, while a bad first impression could cause a loss of reliance or a weaker, less accurate mental model.

When users have freedom of choice and models are imperfect, designers will not have control of a user's first impression with an intelligent system. Therefore, future research is needed to further investigate approaches to continually direct user attention to system strengths and weaknesses throughout user-system interactions. The strength of first impressions motivates the need for improved methods of intervention to combat automation bias, help users develop an accurate understanding of AI capabilities, and develop an appropriate level of trust in intelligent systems.

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] Norman H Anderson and Alfred A Barrios. 1961. Primacy effects in personality impression formation. *The Journal of Abnormal and Social Psychology* 63, 2 (1961), 346.

[3] Hans Brombacher, Dennis Arts, Carl Megens, and Steven Vos. 2019. Stimulight: exploring social

interaction to reduce physical inactivity among office workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, LBW0136.

[4] 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.

[5] Eva Fourakis and Jeremy Cone. 2019. Matters Order: The Role of Information Order on Implicit Impression Formation. *Social Psychological and Personality Science* (2019), 1948550619843930.

[6] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: empirical results assessing influencing factors. *International journal of medical informatics* 83, 5 (2014), 368–375.

[7] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Chau. 2020. Summit: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations. *IEEE Transactions on Visualization and Computer Graphics* 26 (2020). Issue 1.

[8] Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 1069–1078.

[9] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. 2019. Meaningful explanations of Black Box AI decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 9780–9784.

[10] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics* 1 (2013), 25–36.

[11] Arnaud Rey, Kévin Le Goff, Marlène Abadie, and Pierre Courrieu. 2019. The primacy order effect in complex decision making. *Psychological Research* (2019), 1–10.

[12] 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*. ACM, 1135–1144.

[13] 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*.

[14] Jung P Shim, Merrill Warkentin, James F Courtney, Daniel J Power, Ramesh Sharda, and Christer Carlsson. 2002. Past, present, and future of decision support technology. *Decision support systems* 33, 2 (2002), 111–126.

[15] Jessica Sullivan. 2019. The primacy effect in impression formation: Some replications and extensions. *Social Psychological and Personality Science* 10, 4 (2019), 432–439.

[16] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. 2008. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology* 18, 11 (2008), 1473.

[17] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *science* 185, 4157 (1974), 1124–1131.

[18] Emily Wall, Leslie M Blaha, Celeste Lyn Paul, Kristin Cook, and Alex Endert. 2018. Four perspectives on human bias in visual analytics. In *Cognitive biases in visualizations*. Springer, 29–42.

[19] 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*. ACM, 601.

[20] 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)*. IEEE, 435–441.