Examining the Effects of Explainable Hints in AI-Driven Training

Tory Farmer
Washington University in St. Louis
St. Louis, MO, USA
toryfarmer@wustl.edu

Chien-Ju Ho
Washington University in St. Louis
St. Louis, MO, USA
chienju.ho@wustl.edu

ABSTRACT

AI-driven training systems have gained popularity across various disciplines, including disaster response, pilot training, and medical education. Many such systems use "hints," leveraging AI to identify key decision points and suggest actions to aid trainees. Some hint-based AI-driven training systems have been shown to improve trainee performance in deployment when conditions are similar to those in training. However, when conditions in deployment substantially differ from those in training, trainees often fail to generalize the hints to unseen scenarios, resulting in a decrease in performance. Meanwhile, the recent development of explainable AI has provided opportunities to address the generalization challenge by providing explanations to improve humans' understanding of AI hints. In this work, we explore the effect of providing explanations alongside hints in AI-driven training in a simple navigation task. As an exploratory investigation, we utilize large language models (LLMs) to generate explanations about which features of the task led to the given hint. Our preliminary results suggest that, in our simple navigation task with LLM-generated explanations, while providing explainable hints improves trainee performance in environments similar to training, it promotes over-reliance on the AI-provided hints. This results in decreased performance in environments unseen during training. Future work would include examining other mechanisms of explanation generation and investigating the effects in other tasks.

KEYWORDS

Explainable AI, Training, Hint Systems, Reliance

ACM Reference Format:
Tory Farmer and Chien-Ju Ho. 2024. Examining the Effects of Explainable Hints in AI-Driven Training. In *CI '24: ACM Collective Intelligence. June 26-29, 2024, Boston, MA.* ACM, Boston, MA, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As AI becomes increasingly powerful, it offers unprecedented opportunities to enhance human decision-making capabilities, particularly in complex and challenging environments. One common approach for AI to augment human intelligence is to provide suggestions to human decision-makers in the form of *hints*, where a hint is implemented as a suggestion of what action to take at a given time point. In particular, prior works have shown that it is possible to leverage AI techniques to identify effective hints to provide to human decision-makers [3, 7, 18].

While leveraging AI hints to assist humans is promising, offering these hints in deployment is not always practical. It can become too cumbersome or distracting, especially in situations that require focused attention from decision-makers. For example, a pilot evading enemy fire may not be able to process advice from an AI system in deployment due to the need to concentrate on immediate threats. In such scenarios, an alternative is to utilize AI hints during training rather than in deployment, helping trainees learn to avoid potential mistakes in deployment. However, similar to the sim-to-real challenge in robotics [19], there is often an unavoidable gap between the training environments and deployment environments. To effectively utilize AI hints for training, one major challenge is to ensure that humans can *generalize* what they have learned to environments they have not encountered during training.

To achieve generalization, human decision-makers need to understand the AI hints and selectively apply the concepts learned in training to tasks during deployment, as directly mimicking training actions and decisions might result in negative outcomes in previously unseen environments. Given this goal is similar to the recent development of explainable AI [17], in this work, we explore whether providing explanations alongside AI hints during training can help trainees to generalize the hints from training to deployment. In particular, during training time, in addition to providing trainees with AI hints, suggestions on what actions to take at given time points, we also provide explanations that suggest why the hints are generated.

As an exploratory study, we conducted experiments on the Mouselab game developed by Callaway [7]. We leverage their methodologies in generating AI hints. To generate explanations, we utilized recent advancements in large language models (LLMs). We provided ChatGPT with the task and the corresponding hint, asking it to generate explanations about which features of the task led to the given hint. Our results demonstrate that, in our experimental setting, providing explainable hints improves trainee performance in environments similar to training. However, it also promotes over-reliance on the AI-provided hints, resulting in decreased performance in environments not encountered during training. While our preliminary results are discouraging, they highlight the impact of explanations on trainee performance in AI-driven training systems. Moreover, our findings align with existing literature on explainable AI, which often shows that providing explanations can lead to over-reliance on AI outputs [5, 6, 16]. This suggests a potential direction for future research: leveraging insights from the explainable AI literature to mitigate over-reliance and enhance the generalization abilities of trainees.

CI'24, June 2024, Boston, MA, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 RELATED WORK

Our work builds on the work by Callaway [7], who used *metacognitive feedback* (i.e., hints aimed at improving a user's decision-making rather than simply judging their decisions) to enhance users' planning strategies in a simple game. One of the experiments conducted by Callaway [7] involved placing users in a new environment after training (with hints) in a different one, which caused these trainees to perform worse than those in the control group. In other words, the trainees were learning how to play a single specific version of the game and seemingly failed to generalize their new knowledge. In this work, we aim to explore whether this phenomenon could be mitigated by offering explainable hints during training.

In the context of AI-assisted decision-making, using explainable AI has been shown to improve human trust [8], and in some cases, performance [11, 14], across a variety of environments [12]. This improvement is partially due to increased engagement with the task [1]. However, explainability can also lead participants to over-apply the AI's advice when it is not relevant, especially when the hints presented are convincingly wrong [4], wrong too often [15], or too cognitively expensive to engage with [16].

Another related line of work is in the sutudy of far transfer [2], where we look to see if knowledge is transferred from training in one environment to performance in a related, albeit different (in terms of optimal decision-making policy) environment. In this light, hints have been shown to be effective at assisting transferring knowledge between contexts, but only when they assist individuals in making connections between them [2]. However, when viewed as a task switch, hints may very likely cause an over-application of ideas from one task to another, in line with findings on over-reliance [13]. Generally speaking, there are conflicting theories in the literature on how explainable AI would impact knowledge transfers between tasks. Meanwhile, there is relatively limited empirical work to measure the transfers in the presence of explainable AI, which is what our paper aims to explore.

3 EXPERIMENTS AND RESULTS

3.1 Experiment Setup

Our experiments extend the Mouselab game by Callaway [7]. In the Mouselab game, users are ask to control a spider to navigate across the map to collect rewards, as demonstrated. The rewards associated with each node are initially hidden, but users can choose to incur a cost to reveal the reward associated with a given node. The user's planning strategy is to figure out ways to reveal the rewards and ways to navigate the map to maximize the total collected rewards.

At the beginning of the game, the player is placed at the center node. Each non-starting node offers a reward (in coins) indicated by the number on the node, with all rewards initially hidden to the player. At each time step, the player can choose to reveal the reward of a node, incurring a cost of 2 coins or choose to move the spider towards one of the allowable direction. When the spider pass through a node, it collects the associated rewards. When the spider reaches the leaf node, the game ends. The goal of the player is to maximize the total collected rewards, and the player can receive a bonus based on the rewards they collected throughout the game. A dissection of the game interface can be found in Figure 1.

In the training environment, node reward variance increases as a function of distance from the center, i.e., the rewards for the nodes close to the center are close to 0, while the rewards for the nodes far away from the center are larger in magnitude. Therefore, during training, the AI hints would lead users to reveal the rewards for the nodes far away from the center first to identify some high-reward nodes, and then navigate to those nodes to collect the rewards. The LLM-generated explanations provide additional information about the variance of reward distribution not included in hints, e.g., it explains that the reason the AI hints lead users to reveal the nodes far away from the center first is the high-variance property.

To create differences between the training and deployment environments, we vary the reward distributions of the nodes in the deployment environments. Specifically, we separate the deployment environments into "similar-to-training" and "dissimilar-to-training" environments. The "similar-to-training" environments are generated in the same way as the training environments. However, in the "dissimilar-to-training" environments, the nodes closer to the center exhibit higher variance in realized rewards, while the nodes farther away from the center exhibit lower variance in realized rewards. Since participants will not receive AI hints and/or explanations during deployment, we conjecture that participants who only receive AI hints are more likely to continue adopting incorrect strategies for a longer period. In contrast, participants who receive explanations alongside AI hints are likely to adjust more quickly to the change in reward distribution during deployment.

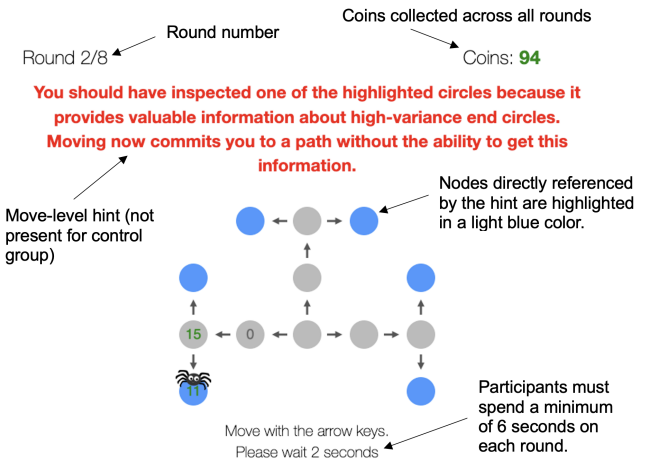

Figure 1: A round of the Mouselab game with explained hint.

For the implementation of AI hints, we leverage the same methodology as Callaway [7], which uses reinforcement learning to determine the optimal policy and generate AI hints based on it. For the generation of explanations, we provide the environment description and the actions from the AI hint to GPT-4 to generate explanations in an offline manner.

3.2 Experiment Procedure

We conducted our experiments using the above experimental setup. Participants in each experiment were placed into one of three treatments: Control, Unexplained Hints, and Explained Hints. The treatments differed in the type of hints participants received during the training phase. All participants experienced the same experimental structure. After agreeing to the informed consent and reading through the instructions, participants proceeded through a training block followed by two deployment blocks. Each block consisted of eight rounds. In the training block and the first deployment block, participants completed tasks in environments similar to the training environments. In the second deployment block, participants completed tasks in "dissimilar-to-training" environments as described above. Before the start of the second deployment block, participants were informed that the environment would change, but they were not provided with specific information on how it would change or what the new optimal strategy would be.

We conducted our experiments by recruiting participants from Prolific. A total of 300 participants were recruited for the experiments, which were approved by the IRB at our institution. We assessed the performance of participants based on the difference between their achieved reward and the maximum possible reward for each round. We refer to this difference as "loss," where a lower loss equates to superior performance.

3.3 Experiment Results

The results are shown in Figure 2. We first found that providing AI hints AI hints increase the performance of participants both during training and during deployment when the deployment environments are similar to those in training. Moreover, when providing explanations alongside AI hints, this performance gain further increases, demonstrating that explanations help humans understand AI hints and improve their performance.

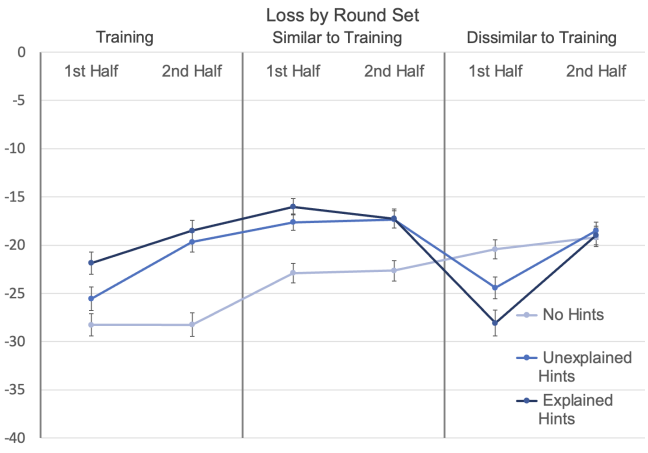

Figure 2: The average loss for each of the three treatments over training and deployment. The loss is defined as the the performance of the participants minus the performance of the performance of the optimal strategy. The error bars represent standard errors.

However, when the deployment environments are different from the training environments, participants receiving AI hints during training significantly underperform participants not receiving AI hints, although these effects dissipate over time (as shown in the second half of the environments in the dissimilar-to-training block). This observation replicates the work by Callaway [7]. Probably more surprisingly, for participants receiving explanations alongside AI hints during training, the performance drop is even more significant than the participants receiving only AI hints during training. This result indicates that participants might have developed over-reliance on the AI hints when provided explanations. Even when the explanations are supposed to help them understand the reasoning of the hints, participants do not internalize the explanations and seemingly developed more reliance on them in their decision-making.

4 CONCLUSION AND DISCUSSION

In this work, we examine the effects of explanations for AI hints in AI-driven training systems. Our results replicate prior findings, showing that providing AI hints during training improves participant performance when the deployment environments are similar to those in training, but it might hurt participant performance when the deployment environments are dissimilar to training. Additionally, when we introduce LLM-generated explanations to AI hints during training, it further enhances participants' performance in similar-to-training deployment environments. However, it also exacerbates the decline in participant performance in dissimilar-to-training environments. These findings suggest that while explanations can be beneficial in familiar contexts, they may lead to over-reliance in novel situations. This highlights the need for further research in the design of AI-driven training systems.

Our work includes a number of limitations, many of which could be grounds for future research. The first limitation concerns the task environments. The game we used is relatively simple and does not cover multiple types of deviations from the training environment. Examining to what extent our results generalize to other environments is a natural next step. Second, in our setting, explanations are always presented alongside AI hints. Literature suggests that showing only the explanations without explicit suggestions on what to do [10] or phrasing the explanation as a question [9] can increase interaction with the hint system and improve performance. Therefore, it would be interesting to explore whether this could be a more effective way of utilizing explanations in AI-driven training. Third, there has been a flourishing line of research in explainable AI, and we have only adopted LLM-generated explanations in our study. It would be important and useful to examine the effects of different ways of implementing explanations. Finally, our results suggest that participants might develop potential over-reliance on AI hints during training when explanations are provided. This finding aligns with recent empirical results that explanations often induce over-reliance in the context of AI-assisted decision making [5, 6, 16]. It is therefore important and interesting to bridge the results and findings of using explanations in the context of AI-driven training and AI-assisted decision making, with the goal of improving our understanding of best leveraging AI to improve human decisions.

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 *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14.

[2] Susan Barnett and Stephen Ceci. 2002. When and Where Do We Apply What We Learn? A Taxonomy for Far Transfer. *Psychological bulletin* 128 (07 2002), 612–37.

[3] Hamsa Bastani, Osbert Bastani, and Wichinpong Park Sinchaisri. 2021. Improving human decision-making with machine learning. *arXiv preprint arXiv:2108.08454* 5 (2021).

[4] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5, CSCW1, Article 188 (apr 2021), 21 pages.

[5] Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th international conference on intelligent user interfaces*. 454–464.

[6] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[7] Frederick Callaway. 2022. Leveraging Artificial Intelligence to Improve People's Planning Strategies. In *PNAS*, Vol. 119.

[8] Cristina Conati, Oswald Barral, Vanessa Putnam, and Lea Rieger. 2020. Toward Personalized XAI: A Case Study in Intelligent Tutoring Systems. arXiv:1912.04464 [cs.AI]

[9] Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 352, 13 pages.

[10] Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces (IUI '22)*. ACM.

[11] Loveleen Gaur, Mehedi Masud, and Noor Jhanjhi. 2022. Explanation-driven HCI Model to Examine the Mini-Mental State for Alzheimer's Disease. (12 2022).

[12] Michael Guevarra, Srijita Das, Christabel Wayllace, Carrie Demmans Epp, Matthew Taylor, and Alan Tay. 2023. Augmenting Flight Training with AI to Efficiently Train Pilots. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 13 (Sep. 2023), 16437–16439. https://doi.org/10.1609/aaai.v37i13.27071

[13] Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. 2010. Control and interference in task switching–a review. *Psychological bulletin* 136 5 (2010), 849–74.

[14] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages.

[15] 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 (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350.

[16] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[17] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8*. Springer, 563–574.

[18] Guanghui Yu and Chien-Ju Ho. 2022. Environment Design for Biased Decision Makers.. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 592–598.

[19] Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. 2020. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *2020 IEEE symposium series on computational intelligence (SSCI)*. IEEE, 737–744.