Adaptive interventions for both accuracy and time in AI-assisted human decision making

Siddharth Swaroop [1] Zana Buçinca [1] Finale Doshi-Velez [1]

Abstract

In settings where users are both time-pressured and need high accuracy, such as doctors working in Emergency Rooms, we want to provide AI assistance that both increases accuracy and reduces time. However, different types of AI assistance have different benefits: some reduce time taken while increasing overreliance on AI, while others do the opposite. We therefore want to adapt what AI assistance we show depending on various properties (of the question and of the user) in order to best trade off our two objectives. We introduce a study where users have to prescribe medicines to aliens, and use it to explore the potential for adapting AI assistance. We find evidence that it is beneficial to adapt our AI assistance depending on the question, leading to good tradeoffs between time taken and accuracy. Future work would consider machine-learning algorithms (such as reinforcement learning) to automatically adapt quickly.

1. Introduction

Artificially intelligent (AI) systems are being used to help humans in many settings make decisions or predictions, ranging from helping doctors in disease diagnosis (Musen et al., 2014) to helping judges make pretrial-release decisions (Green & Chen, 2019). Most studies focus on how different AI assistance types (for example, providing only an AI recommendation, or providing an AI recommendation and explanation) impact the overall accuracy of the human decision-maker, often finding that humans can overrely on an AI prediction (Bussone et al., 2015; Lai & Tan, 2019; Jacobs et al., 2021). Recent work suggests that the cognitive effort induced by an AI assistance may also play a part in the overreliance rate and accuracy achieved (Buçinca et al., 2021). Other studies look at how AI assistance impacts how long a human takes to make a decision, with mixed findings (Arshad et al., 2015; Fogliato et al., 2022).

However, instead of focussing on a single metric, in many settings we may need to consider how AI assistance impacts multiple metrics. For example, a doctor might be under stringent time constraints, such as needing to address a long queue of patients at an Emergency Room, and needs to obtain as high an accuracy as possible given these constraints (Patel et al., 2008; Franklin et al., 2011; Rundo et al., 2020). In this paper, we focus on how AI assistance impacts these two metrics in detail: accuracy and time.

Different types of AI assistance trade off accuracy and time differently, and previous results indicate that no single type leads to both optimal accuracy and minimal time taken. These tradeoffs may also be related to the cognitive effort or cost required (Buçinca et al., 2021; Vasconcelos et al., 2023), and how much humans overrely on the AI prediction. AI assistance types that require more cognitive effort take longer to process, but can lead to higher accuracy. Conversely, AI assistance types requiring less cognitive effort, such as providing an AI recommendation, can lead to overreliance on the AI assistance (Bussone et al., 2015; Lai & Tan, 2019; Jacobs et al., 2021), which may be undesirable (and potentially lowers accuracy). Additionally, when shown an AI assistance type requiring higher cognitive effort, humans with a higher intrinsic motivation to think (their Need for Cognition (NFC) trait) may perform better (Buçinca et al., 2021).

In general, we therefore want to adapt what assistance is shown depending on the question and the human, and we explore this idea in this paper. On easier questions, where humans are less likely to overrely on AI recommendations or where AI accuracy is likely to be higher, we can reduce the cognitive effort required and reduce time without sacrificing accuracy. On harder questions, we may want humans to engage in more depth, increasing time in order to increase accuracy. Additionally, we may have to adapt our AI assistance to different people. For example, humans with higher NFC may see the benefits of increased cognitive effort more, and have higher tolerance to such AI assistances. Humans with a higher skill (for example, more proficient doctors) may not require as much assistance from an AI.

[1] Harvard University, USA. Correspondence to: Siddharth Swaroop <siddharth@seas.harvard.edu>.

AI & HCI Workshop at the 40^{th} International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

In this paper, we introduce a task where users are motivated to do well in both metrics. In our alien prescription task, participants have to prescribe medicine to a series of sick aliens, and have a set time to get through as many aliens as possible, while maintaining a high accuracy. Our pilot studies indicate that different AI assistance types trade off accuracy and time in different ways, and that it would be beneficial to adapt the assistance depending on various properties of the question, such as adapting to the question's difficulty. We also find that, as time progresses in the study, users start making more mistakes, and so we may also want to adapt the AI assistance depending on how much overall time has passed. We also expect each user's NFC and skill to be important to adapt to, but leave this for future work.

2. Related Works

AI-assisted human decision making. Initial studies expected AI+human teams to perform better than either alone (Kamar et al., 2012; Amershi et al., 2019), however, recent studies have found that this is not the case, with accuracy of the team usually worse than AI-only accuracy (Bussone et al., 2015; Lai & Tan, 2019; Green & Chen, 2019; Bansal et al., 2021). This may be because humans overrely on AI predictions, making mistakes by agreeing with a wrong AI prediction (even when the human may not have made the mistake on their own), instead of achieving complementary performance (Bussone et al., 2015; Lai & Tan, 2019; Jacobs et al., 2021). As a way to combat this, Buçinca et al. (2021) introduced cognitive forcing functions as interaction design interventions to reduce overreliance on AI. They showed that the update condition, in which participants are asked to make a decision on their own first before seeing an AI recommendation, reduced overreliance. But the update condition may also reduce appropriate reliance, as experts may pay less attention to the recommendation after spending effort and time to make the decision unassisted (Fogliato et al., 2022). We explore the update condition in our work, finding it can increase accuracy and reduce overreliance (although not eliminate it entirely).

Adaptive interventions. A few recent studies have considered adapting the AI assistance shown to users. Noti & Chen (2022) train a classifier on previous data to adaptively show AI recommendations. They find that they can increase AI+human performance by showing AI recommendations only on questions that the AI is more likely to be right. Ma et al. (2023) also find similar results in their setting. Bhatt et al. (2023) consider adapting the form of AI assistance shown to different users' preferences, using contextual bandits to trade off accuracy against the cost of assistance. Overall, we believe these results show that adaptive interventions are a promising research direction, and we consider their potential to trade off accuracy and time.

Accuracy and time tradeoff. To the best of our knowledge, no prior work has focussed explicitly on the tradeoff between accuracy and time in AI-assisted decision-making. Multiple studies, however, report response times of participants when shown different conditions or interventions. However, the results present mixed empirical evidence. Some studies find that people spend more time on instances that they perceive as more difficult inherently (Arshad et al., 2015; Levy et al., 2021), but this additional time spent does not translate to increased accuracy. We also find this in the absence of any AI assistance. For clinical annotations, Levy et al. (2021) found that despite additional time spent on instances with incorrect AI recommendations, accuracy was lower compared to instances with correct recommendations. Fogliato et al. (2022) found that time spent on the task did not differ among standard and update conditions, while we find that response time does increase in the update condition in our setting.

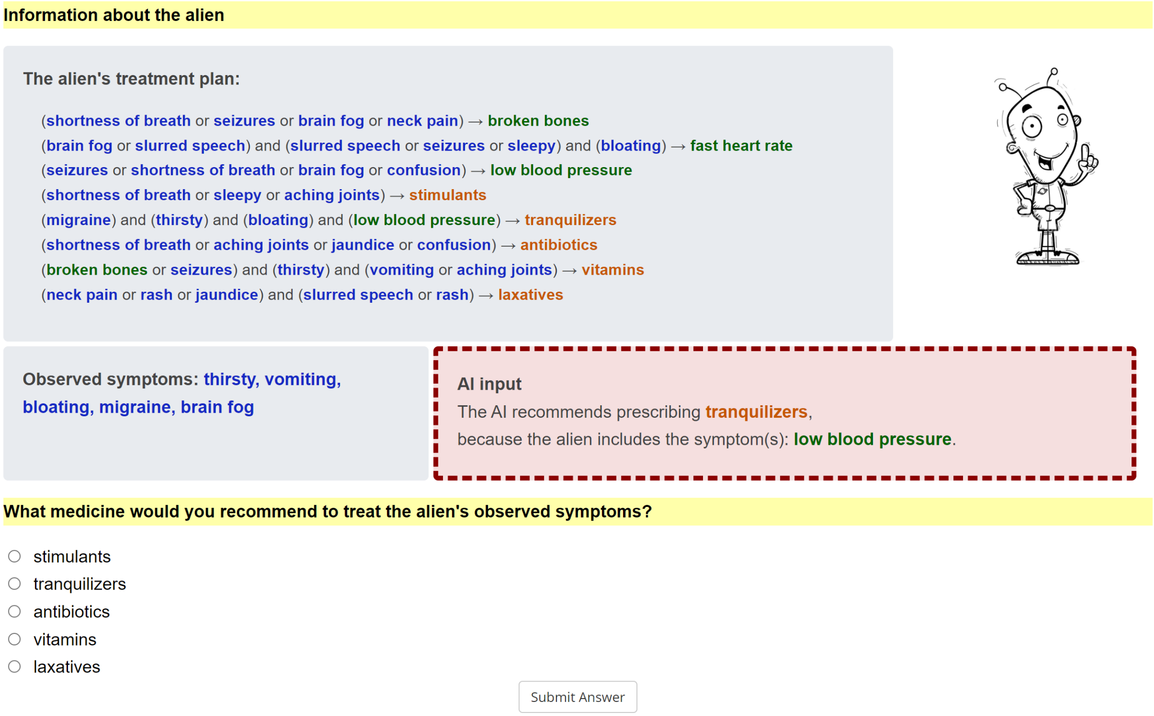
3. Experiments

In this section, we describe the alien prescription task, the various design choices we made, and the procedures for the pilot studies we ran.

Alien prescription task design. We designed a task where users are asked to prescribe medicines to sick aliens, which we base on a previous work (Lage et al., 2019). Participants were shown a series of sick aliens for a fixed time of 15-20 minutes, corresponding to their 'medical shift', and asked to prescribe a single medicine to each alien. By asking participants to act like doctors, and by emphasizing the importance of treating patients correctly, we aimed to motivate participants to obtain a high accuracy, while getting through as many sick patients as possible during their medical shift.

Figure 1 shows an example of a single alien task. Based on observed symptoms and the 'treatment plan' (which is a set of decision set rules unique to each alien), participants must decide a single medicine to give the alien. We chose decision sets as they are relatively easy for humans to parse (Lakkaraju et al., 2016). When we provide an AI assistance, we show it in a red box, as shown in Figure 1. This box can provide both an AI recommendation and explanation (explanations are always an intermediate symptom that lead to the recommended medicine), and is provided before or after the participant's initial decision.

We expanded on the setup in Lage et al. (2019) in three ways. First, we always introduced intermediate symptoms to the task, which require participants to perform additional computation steps, and worked well as the explanation of an AI's recommendation. Second, we also allowed two possible correct medicines per alien. We defined the better medicine to be one that addressed more of the observed



Information about the alien

The alien's treatment plan:

(shortness of breath or seizures or brain fog or neck pain) → broken bones
(brain fog or slurred speech) and (slurred speech or seizures or sleepy) and (bloating) → fast heart rate
(seizures or shortness of breath or brain fog or confusion) → low blood pressure
(shortness of breath or sleepy or aching joints) → stimulants
(migraine) and (thirsty) and (bloating) and (low blood pressure) → tranquilizers
(shortness of breath or aching joints or jaundice or confusion) → antibiotics
(broken bones or seizures) and (thirsty) and (vomiting or aching joints) → vitamins
(neck pain or rash or jaundice) and (slurred speech or rash) → laxatives

Observed symptoms: thirsty, vomiting, bloating, migraine, brain fog

AI input
The AI recommends prescribing tranquilizers, because the alien includes the symptom(s): low blood pressure.

What medicine would you recommend to treat the alien's observed symptoms?

○ stimulants
○ tranquilizers
○ antibiotics
○ vitamins
○ laxatives

Figure 1. The alien prescription task, where participants must prescribe a single medicine. The information about the alien includes the alien's unique treatment plan (a set of rules) and the alien's observed symptoms. Participants have to use these observed symptoms and rules to prescribe a single medicine, such that only the observed symptoms and any potential intermediate (green) symptoms are used, and no other unobserved symptoms. When an AI assistance is shown, it is shown in a red box, like in this example. Here, the AI recommendation is the best possible (tranquilizers uses the most observed symptoms). Vitamins is also a correct medicine, but is suboptimal as it uses fewer observed symptoms. All other medicines are incorrect.

symptoms. Having a suboptimal medicine helps us to better analyze the role of overreliance on AI recommendations: suboptimal medicines are easily verified to be correct, and so participants can overrely on them more easily than overrelying on a wrong recommendation.

Third, we introduced two different difficulties of questions: easy and hard. We designed these such that easy questions require less cognitive effort for a human to find the best medicine, while hard questions require more computation. We ensured that both easy and hard questions superficially look very similar to a human, by having a similar length of lines, number of lines, and other visual aspects. Further discussion and examples are in Appendix B.

Interventions. We consider three AI assistance types.

1. *No-AI*: Do not provide any AI assistance.
2. *AI before*: An AI recommendation and explanation is provided to the participant along with the question, before the participant makes any decision.
3. *AI after (/update)*: The participant makes an initial decision without any AI assistance. They are then provided with an AI recommendation and explanation, and allowed to change their initial answer.

In Appendix A, we also look at the effect of providing only the AI explanation (no AI recommendation) both before and after the user's initial decision. Our initial results indicate that this explanation-only assistance does not help participants, who often perform worse than if no AI assistance had been given. We believe this is due to the form of the explanation, and different explanations might lead to different results. Please see Appendix A for more discussion.

For most of our results, we had a timer shown on the screen, indicating the time remaining for participants to answer questions (the length of their 'medical shift'). We also examine how participants perform when no timer is shown in Figure 2.

Procedure. Before starting the main part of the study, participants had to accept a consent form, read instructions, and successfully complete three practice questions (for which they had two attempts, similar to Lage et al. (2019)). In the initial pilot studies, the participants then had 15 minutes to answer as many questions as possible. These pilot studies were split into two halves (the order of the halves was randomized): one with a particular AI assistance type, and one with no-AI. This allowed us to both change the difficulty of the question (as measured by performance without AI assistance), and measure the effects of a specific AI assistance. In our first pilot studies, we found that participants found the task too easy, with many participants spending less than 20 seconds per question while achieving 100% accuracy. We increased the difficulty (such as by increasing the number and length of lines, and increasing the number of observed symptoms (Lage et al., 2019)) until participants took about one minute per easy question. Approximately half of the questions participants saw were easy questions, and the other half hard questions.

Once we fixed the difficulty of the questions, later pilot studies were 20 minutes long, with AI assistance type randomly assigned to each question. By increasing the length of the study, we expected participants to get more tired during the study (and less willing to cognitively engage with questions). By randomly assigning AI assistance type to questions, we make the pilot studies more realistic, as eventually the AI assistance type will be adaptively assigned to each question. Our results in Figure 2 are from these later pilot studies.

After the study, participants were shown a final screen where they were asked what their strategy was (and if it changed when there was an AI input), and for any other feedback.

Participants. We ran six pilot studies on Prolific, with 20 participants each. Only English speakers were allowed to participate. In each study, we remove 3-7 participants from analysis, as they either failed the practice questions or let the timer run out without answering questions.

Participants were paid at a rate of $12 per hour ($6 for the 30 minute studies, and $7 for the 35 minute studies; these times include 15 minutes for reading instructions and completing practice questions). We also incentivized participant performance by providing a bonus $3 reward to the top-performing participant in each study. If the participants failed the practice questions twice, the study ended early, with a smaller pay of $2. In general, we found that participants seemed to engage positively with the study, with many commenting that they had tried their best to treat their sick alien patients, and that they found the study was well-designed.

Design and analysis. We report three metrics.

1. *Accuracy*: if participants chose the *best* medicine for the alien, we gave them a score of 1, a *suboptimal* (but correct) medicine has a score of 0.5, and a *wrong* medicine has a score of 0. We calculate the average accuracy over questions for each participant, and report mean and standard error across participants.
2. *Response time*: we measure how long each participant takes to answer questions. We report the mean and standard error across participants.
3. *Overreliance*: we define overreliance to be the proportion of times a participant gave the same answer as the AI when the AI was wrong or suboptimal (Buçinca et al., 2021; Vasconcelos et al., 2023).
4. *Underreliance*: we define underreliance to be the proportion of times a participant gave a non-optimal answer when the AI was optimal.

Figure 2 also shows how participant accuracy and response time changes during the course of the study. To plot this figure, we find each participant's most recently answered question, and plot mean and standard error across the appropriate metric (accuracy or response time). We repeat this at equally-spaced intervals over the course of the study.

When we show an AI assistance, there is 60% chance that the AI recommends the best medicine, 30% chance that the AI is suboptimal, and 10% chance that the AI is wrong. The explanation is chosen such that it is faithful to the (correct or incorrect) AI recommendation. Overall, this typically gives an average AI-only accuracy of 0.79±0.03 (mean and standard error across participants). We purposefully ensured that the AI-only accuracy is similar to human-only accuracy.

4. Results

Our results show that, in our alien prescription task, there is potential for adapting the type of AI assistance depending on the question, in order to achieve good time-accuracy tradeoffs. We leave a detailed look at the effect of cognitive effort and NFC for future work.

The results we present are based on pilot studies, with 15-20 participants per study. The trends we see are promising for a future, larger study that we plan to run. Due to the small sample sizes, we do not report p-values.

Difficulty	Avg acc	Avg time (s)
All	0.80±0.05	66±7
Easy	0.91±0.02	60±7
Hard	0.69±0.08	76±8

Table 1. Mean and standard error of accuracy and response time on the No-AI condition ($n = 14$ participants). Humans achieve an accuracy of 80% on average, taking 66 seconds per question. Humans have higher accuracy and are quicker on 'easy' questions.

Performance without any AI assistance. We first look at participant performance with the No-AI condition, summarized in Table 1. We see that average accuracy is 0.80, and the average response time is 66 seconds. We can also see the difference between the two question difficulties, easy and hard. On easy questions, participants are marginally quicker (60 seconds) and have higher accuracy (0.91). On hard questions, participants are slower (76 seconds) and have lower accuracy (0.69).

We also note that the standard deviation across participants is large, as some participants are quicker and/or better at the task than others. The standard deviation in accuracy is 0.18, and in response time is 26 seconds (note that we report standard error in Table 1, not standard deviation). Therefore, when comparing to different AI assistance types, we look at how each AI assistance type impacts each participant separately (by comparing metrics to the No-AI condition), and then average across participants.

Question difficulty	AI condition	Change in avg acc	Change in avg time (s)
All	AI before	-0.005±0.03	-13±3
	AI after	0.09±0.03	9±1
Easy	AI before	-0.04±0.04	-11±4
	AI after	0.02±0.03	8±1
Hard	AI before	0.007±0.06	-12±6
	AI after	0.17±0.07	9±2

Table 2. Effect of AI assistance types, measured as within-participant differences to the No-AI condition (mean and standard error). The AI before condition ($n = 17$ participants) saves time without significantly impacting accuracy, and can be used on easy questions. The AI after condition ($n = 14$ participants) increases response time, but increases accuracy on hard questions.

The AI before condition reduces time taken, but increases overreliance. We see in Table 2 that the AI before condition does not impact average accuracy significantly, but does reduce time taken to answer questions. This is the case in both easy and hard questions. Participants' overreliance rate is 48±9% in this case, which is high. Conversely, participants' underreliance rate is 8±3%, which is very low, showing they usually trust the AI recommendation.

The AI after condition increases time taken, but also increases accuracy on hard questions. The AI after condition, on average, increases time taken (by 9 seconds), and increases accuracy. This is because participants are able to spot mistakes they made, and correct them, without necessarily overrelying on the AI recommendation. In fact, overreliance rate is significantly lower than with the AI before condition, at 14±8%, while underreliance rate is similarly low at 12±4%. This indicates that participants use the AI input to cognitively engage more with the question, similar to results in previous studies (Buçinca et al., 2021; Vasconcelos et al., 2023). In fact, we found that participants never changed their answer to match the AI's recommendation when the AI was suboptimal or wrong, but did sometimes change their answer to agree with the AI recommendation when the AI was right. When we analyze these results in terms of easy and hard questions, we find that participant accuracy does not significantly increase on easy questions (likely as accuracy is already high). On hard questions however, participant accuracy does increase. This indicates that the AI after condition is particularly beneficial when shown on hard questions.

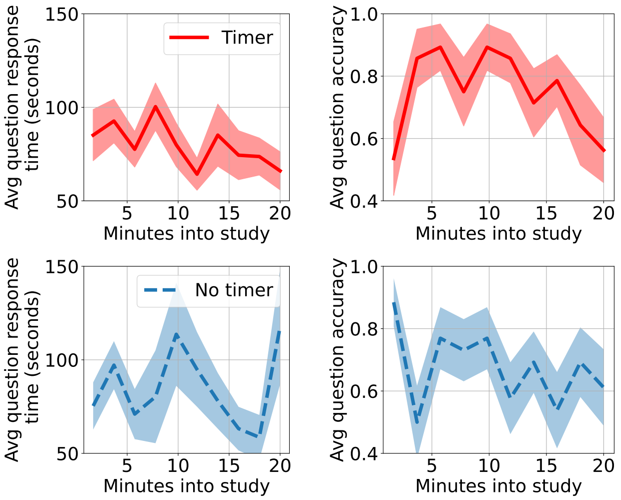


Figure 2. Plots of response time and accuracy during the course of the study. Top row: when a timer is shown to participants ($n = 14$ participants), they maintain a fast pace of answering questions (top left), but accuracy reduces later in the study (top right). Bottom row: When no timer is shown to participants ($n = 13$ participants), they maintain a constant accuracy (bottom right), but average response time is high (bottom left).

When a timer is shown, users maintain a fast pace of answering questions, but accuracy reduces later in the study. We next look at how participants' accuracy and response time changes during the course of the 20-minute studies. Figure 2 (top) shows results for when a timer is shown to the participants. We see that users maintain a fast pace of answering questions throughout the study (perhaps even getting marginally faster due to the time pressure). However, accuracy tends to decrease slowly. In the first 5 minutes of the study, participants are likely still learning how best to answer questions, which leads to the initial increase in accuracy. Overall, we see that the time pressure due to a timer leads to a reduction in accuracy over the duration of the study, as participants appear to feel pressured into answering questions as quickly as possible. We hypothesize that this reduction may also occur because participants are getting tired during the course of the study, and are no longer willing to expend as much cognitive effort.

When there is no timer, users maintain a constant accuracy during the course of the study. In Figure 2 (bottom), we see that when no timer is shown on the screen, participants maintain a constant accuracy. Figure 2 (bottom left) shows that question response time changes a lot during the study, and we believe this may be because of noise due to the small sample size. However, overall, response times are larger than when a timer is shown, and accuracy is lower.

5. Discussion

Our results indicate that cleverly choosing AI assistance type can lead to good tradeoffs between time taken and overall accuracy. For instance, providing AI assistance before a user's initial decision can save time when the AI is likely to be right, particularly on easy questions. Providing AI assistance after a user's initial decision increases response time, but causes users to engage more with the question, increasing accuracy (without increasing overreliance). We can do this when the user's initial decision disagrees with the AI's recommendation. When time-pressured (with a timer shown on the screen), users drop in accuracy while maintaining a constant response time per question. We could try to slow down users in order to increase their accuracy.

Using the results from these pilot studies, we intend to show the benefits of adapting AI assistance type on a larger study.

We also expect that we should adapt to different humans based on their intrinsic motivation to think (their NFC trait), as well as potentially their overall skill level on the task. For example, humans with a higher NFC may react better to AI assistance types that prompt them to think more (Buçinca et al., 2021). They may also not make as many mistakes later on in the study. We leave a detailed analysis of this for future work. In our current pilot studies, any signal regarding participants with different NFC is too small to comment on. We hope that a larger-scale study would show any signal more.

Eventually, in a more general setting, we may want to use reinforcement learning to adaptively choose the AI assistance type depending on properties of the question (such as difficulty) and the human (such as NFC).

We note that our study is conducted in a low-stakes environment, and future work should look at how we can adaptively choose AI assistance type in a high-stakes environment too. Our results would not generalize if there are sufficiently different pressures and stakes in such settings. However, it should still be beneficial to adapt the AI assistance type.

Future work could also look at the form of the explanation shown. In our case, the explanation was an intermediate symptom, and is verifiable by the participants. This verifiability might make it easier for participants to avoid overrelying on the AI assistance, especially in the AI after condition (Fok & Weld, 2023). Other forms of explanation may not be easily verifiable, and may lead to different results.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. IIS-2107391. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Quinn, K. I., Teevan, J., Kikin-Gil, R., and Horvitz, E. Guidelines for human-ai interaction. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

Arshad, S. Z., Zhou, J., Bridon, C., Chen, F., and Wang, Y. Investigating user confidence for uncertainty presentation in predictive decision making. In *Proceedings of the annual meeting of the Australian special interest group for computer human interaction*, pp. 352–360, 2015.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2021.

Bhatt, U., Chen, V., Collins, K. M., Kamalaruban, P., Kallina, E., Weller, A., and Talwalkar, A. Learning personalized decision support policies. *arXiv preprint arXiv:2304.06701*, 2023.

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. 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):1–21, 2021.

Bussone, A., Stumpf, S., and O'Sullivan, D. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pp. 160–169, 2015. doi: 10.1109/ICHI.2015.26.

Fogliato, R., Chappidi, S., Lungren, M., Fisher, P., Wilson, D., Fitzke, M., Parkinson, M., Horvitz, E., Inkpen, K., and Nushi, B. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In

2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1362–1374, 2022.

Fok, R. and Weld, D. S. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *arXiv preprint arXiv:2305.07722*, 2023.

Franklin, A., Liu, Y., Li, Z., Nguyen, V., Johnson, T. R., Robinson, D., Okafor, N., King, B., Patel, V. L., and Zhang, J. Opportunistic decision making and complexity in emergency care. *Journal of Biomedical Informatics*, 44(3):469–476, 2011. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2011.04.001. URL `https://www.sciencedirect.com/science/article/pii/S1532046411000657`. Biomedical Complexity and Error.

Green, B. and Chen, Y. The principles and limits of algorithm-in-the-loop decision making. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), nov 2019. doi: 10.1145/3359152. URL `https://doi.org/10.1145/3359152`.

Jacobs, M., F.Pradier, M., McCoy, T., Perlis, R., Doshi velez, F., and Gajos, K. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry*, 11, 02 2021. doi: 10.1038/s41398-021-01224-x.

Kamar, E., Hacker, S., and Horvitz, E. Combining human and machine intelligence in large-scale crowdsourcing. 2012.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S., and Doshi-Velez, F. An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1902.00006*, 2019.

Lai, V. and Tan, C. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 29–38, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287590. URL `https://doi.org/10.1145/3287560.3287590`.

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1675–1684, 2016.

Levy, A., Agrawal, M., Satyanarayan, A., and Sontag, D. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2021.

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., and Ma, X. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. doi: 10.1145/3544548.3581058. URL `https://doi.org/10.1145/3544548.3581058`.

Musen, M., Middleton, B., and Greenes, R. *Clinical decision-support systems*, pp. 643–674. Springer London, United Kingdom, January 2014. ISBN 9781447144731. doi: 10.1007/978-1-4471-4474-8_22.

Noti, G. and Chen, Y. Learning when to advise human decision makers. *arXiv preprint arXiv:2209.13578*, 2022.

Patel, V., Zhang, J., Yoskowitz, N., Green, R., and Sayan, O. Translational cognition for decision support in critical care environments: A review. *Journal of biomedical informatics*, 41:413–31, 07 2008. doi: 10.1016/j.jbi.2008.01.013.

Rundo, L., Pirrone, R., Vitabile, S., Sala, E., and Gambino, O. Recent advances of hci in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of Biomedical Informatics*, 108:103479, 2020. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2020.103479. URL `https://www.sciencedirect.com/science/article/pii/S1532046420301076`.

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., and Krishna, R. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–38, 2023.

A. Providing an AI explanation only (no AI recommendation)

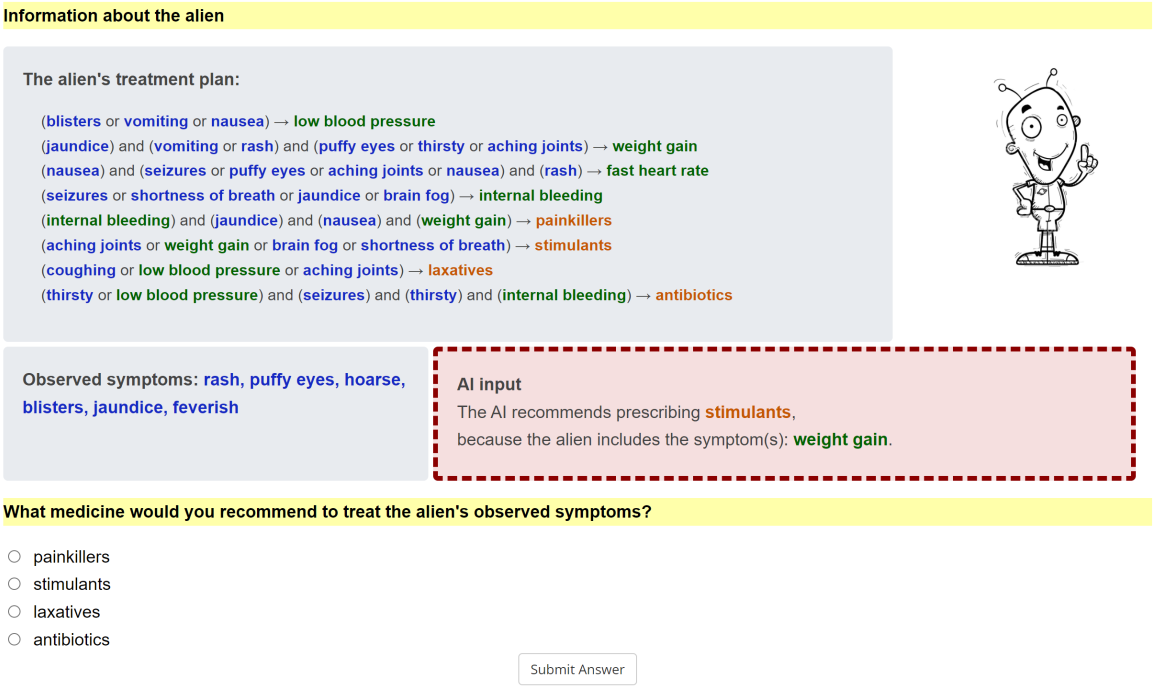
Providing only an ‘explanation’ before the user’s initial decision does not seem to improve accuracy or response time. Accuracy is similar to No-AI accuracy and the AI before condition (which shows both a recommendation and explanation), while response time is faster than No-AI but slower than AI before. However, overreliance rate is better (lower) than the AI before condition. These results hold regardless of question difficulty (easy or hard) or if there is time pressure (timer shown on screen or not). Similar results also hold when the explanation is shown *after* the participant’s initial decision (response time is longer than No-AI, but still shorter than the AI after condition; accuracy is similar to No-AI, but worse than AI after).

Future work could also look at the form of the explanation shown. In our case, the explanation is an intermediate symptom, and is verifiable by the participants. This verifiability might make it easier for participants to avoid overrelying on the AI assistance, especially in the AI after condition (Fok & Weld, 2023). Other forms of explanation may not be easily verifiable, and may lead to different results.

B. Easy and hard questions

We designed two difficulties of questions, such that easy questions require less cognitive effort for a human to find the best medicine, while hard questions require more computation. This is similar to real-world settings where some tasks are more difficult, for example some patients require more analysis to diagnose. We ensured that both easy and hard questions superficially look similar to the participants (similar length of lines, number of lines, and other visual aspects), so that participants cannot easily decide which question will require more cognitive effort before engaging with the question.

Figure 1 shows an example of an easy question. Here, the optimal medicine is the one which uses the most symptoms. Therefore, after finding the optimal medicine, participants only need to confirm that all other medicines use fewer symptoms, and do not need to verify if the symptoms used are observed or not. Figure 3 shows an example of a hard question. Now, the optimal medicine uses very few observed symptoms, and there are many incorrect medicines that use more symptoms. This requires participants to check the other medicines and confirm that they are incorrect, before concluding that the medicine they found is the optimal one. Table 1 shows how participants have higher accuracy and shorter response times on easier questions.



Information about the alien

The alien's treatment plan:

(blisters or vomiting or nausea) → low blood pressure
(jaundice) and (vomiting or rash) and (puffy eyes or thirsty or aching joints) → weight gain
(nausea) and (seizures or puffy eyes or aching joints or nausea) and (rash) → fast heart rate
(seizures or shortness of breath or jaundice or brain fog) → internal bleeding
(internal bleeding) and (jaundice) and (nausea) and (weight gain) → painkillers
(aching joints or weight gain or brain fog or shortness of breath) → stimulants
(coughing or low blood pressure or aching joints) → laxatives
(thirsty or low blood pressure) and (seizures) and (thirsty) and (internal bleeding) → antibiotics

Observed symptoms: rash, puffy eyes, hoarse, blisters, jaundice, feverish

AI input
The AI recommends prescribing **stimulants**,
because the alien includes the symptom(s): **weight gain**.

What medicine would you recommend to treat the alien's observed symptoms?

○ painkillers
○ stimulants
○ laxatives
○ antibiotics

Figure 3. An example of a hard question (see Figure 1 for an example of an easy question). Although this hard question looks superficially similar to an easy question, it requires more cognitive effort to solve. This is because there are many medicines that use many symptoms, and the optimal medicine uses only a few symptoms. This requires participants to manually check the other medicines and confirm that they are incorrect (because they use unobserved symptoms), before concluding that the medicine they found is the optimal one. Note that here, the AI recommendation is again optimal: stimulants uses the most observed symptoms.