Personalising AI assistance based on overreliance rate in AI-assisted decision making

Siddharth Swaroop
Harvard University
Cambridge, Massachusetts, USA
siddharth@seas.harvard.edu

Zana Buçinca
Harvard University
Cambridge, Massachusetts, USA
zbucinca@seas.harvard.edu

Krzysztof Z. Gajos
Harvard University
Cambridge, Massachusetts, USA
kgajos@g.harvard.edu

Finale Doshi-Velez
Harvard University
Cambridge, Massachusetts, USA
finale@seas.harvard.edu

Abstract

Personalising decision-making assistance to different users and tasks can improve human-AI team performance, such as by appropriately impacting reliance on AI assistance. However, people are different in many ways, with many hidden qualities, and adapting AI assistance to these hidden qualities is difficult. In this work, we consider a hidden quality previously identified as important: overreliance on AI assistance. We would like to (i) quickly determine the value of this hidden quality, and (ii) personalise AI assistance based on this value. In our first study, we introduce a few probe questions (where we know the true answer) to determine if a user is an overrelier or not, finding that correctly-chosen probe questions work well. In our second study, we improve human-AI team performance, personalising AI assistance based on users' overreliance quality. Exploratory analysis indicates that people learn different strategies of using AI assistance depending on what AI assistance they saw previously, indicating that we may need to take this into account when designing adaptive AI assistance. We hope that future work will continue exploring how to infer and personalise to other important hidden qualities.

CCS Concepts

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

Keywords

AI-assisted decision-making, time pressure, overreliance, reinforcement learning, adaptive AI, human-centered AI, explainable AI, human-AI interaction, decision support systems

ACM Reference Format:

Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2025. Personalising AI assistance based on overreliance rate in AI-assisted decision making. In *30th International Conference on Intelligent User Interfaces (IUI '25), March 24–27, 2025, Cagliari, Italy*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3708359.3712128

IUI '25, March 24–27, 2025, Cagliari, Italy

ACM ISBN 979-8-4007-1306-4/25/03
https://doi.org/10.1145/3708359.3712128

1 Introduction

In AI-assisted human decision making, our goal is to design AI assistances that complement the human's performance. However, using a fixed type of AI assistance for all people does not lead to complementarity [3, 6, 10, 19, 36, 51]. Recent work has found that adapting the AI assistance to people can help overcome this issue: different people benefit from different types of AI assistance [5, 8, 30, 32]. However, it is not always clear what human qualities we should adapt our AI assistance to. For example, different people can engage different amounts with the task, trust AI differently, and learn about the task differently. Some of these differences might be fixed traits of people and potentially measurable through questionnaires, while others might be specific to the person's daily circumstances and the nature of the task. Examples may include a person's skill at the task or their propensity to rely on the specific AI assistance for the specific task.

We want to adapt to these changing person *qualities*, but they are often hidden or latent: we need to infer these qualities from observing the person in real-time. It is especially difficult to quickly personalise to different people based on hidden qualities, because any real-time signal is noisy, and signal from questionnaires are not very predictive (for example, personality traits asking about trust in AI have not been found to be predictive of a person's reliance on AI [25, 38, 46]).

In this paper, we consider how to quickly adapt AI assistance to one specific hidden quality: overreliance rate (which is how often people rely on an incorrect AI recommendation [7, 46, 47]). Previous work [46] found that this is an important hidden quality of decision-makers: different AI assistances are beneficial for people with different overreliance quality. For example, people with low overreliance tend to be slower and achieve higher accuracy, even achieving human-AI complementarity, unlike people with high overreliance. Additionally, we expect that a person's tendency to overrely will depend on the specific task setting, meaning we must re-learn how to adapt to overreliance quality every time a person approaches a new task.

We use a study where a participant has to answer a series of logic puzzles. Participants are told that they are a doctor that has to treat as many alien patients as possible within a 20-minute shift,

and there is a timer on screen to pressure them to answer quickly. Participants are either shown no AI assistance, AI-before assistance (where the AI recommendation and explanation are shown before the participant makes an initial decision) or AI-after assistance (where the AI input is shown after the participant makes an initial decision, and they can update their decision). These AI assistance types have been found to affect overreliance rate, accuracy and time taken in different ways, and were used in previous work [7, 46]. We want to show the benefit of personalising even between these three assistance types.

Using previous data [46], we design different assistance policies for overreliers and not-overreliers. We assume we have access to (i) whether the particular question is easy or hard, (ii) the uncertainty of the AI recommendation, and (iii) whether the participant is an overrelier or a not-overrelier. Our policies show that people with different overreliance quality should be shown different AI assistance types. For example, people who overrely more should not be shown an AI recommendation in situations when the AI is uncertain. However, people who overrely less can still benefit from being shown an AI recommendation when the AI is wrong, as they can verify it is wrong and then use this to reach a correct answer quickly.

We conducted two studies in this paper. In our first study (Section 3), we explore how we can quickly learn a person's hidden overreliance quality (if they are an overrelier or a not-overrelier). To do this, we introduce 'probe questions', where we purposefully provide an incorrect AI recommendation to see if the participant overrelies on it or not. These are similar to catch trials, where an incorrect suggestion is shown to a user with the aim of ensuring the user is vigilant and alert [26, 50]. Overall, our findings show that we can use probe questions to quickly estimate a person's hidden overreliance trait: we get >90% accuracy at predicting overreliance quality (their overreliance on subsequent trials) after only two probe questions.

We also try and understand which personality traits correlate with overreliance, and how people's overreliance quality affects their subjective experience and motivation about the task. We find that overreliers have significantly lower Need-for-Cognition trait (people's intrinsic motivation to think [12]), while previous work only found marginal correlations [46]. We also ask participants intrinsic motivation questions (from Self-Determination Theory) after they complete the study [15, 40] to understand whether overreliance quality predicts people's subjective experience of the task. We find that participants who overrely less feel like they put in significantly more effort, feel more pressure, and have higher perceived choice than overreliers. This indicates that these not-overreliers engage with the task more, and feel the time pressure more.

In our second study (Section 4), we explore if we can improve accuracy-time performance by personalising quickly to the hidden overreliance quality. We adapt quickly, showing our personalised policy after just two probe questions instead of half-way through the study. We find that, for overreliers, the personalised policy improves overreliers' accuracy compared to baselines, as expected. For not-overreliers, all policies have similar performance: this group engages with the task, and all forms of AI assistance, more.

In Section 5 we explore how per-question response time is impacted by adapting policies to participants' overreliance quality. We find that our personalised policy speeds up per-question response time on average (or is as quick) compared to baseline policies. Our analysis indicates that when personalising AI assistance, it matters how long participants have spent using previous assistance policies, as this affects how participants use or view the AI input: for example, seeing only AI-before assistance leads to learning a different strategy than if they saw a mix of all AI assistances randomly. This suggests that we could explicitly model a participant's strategy for using AI input, as this may be a more direct indicator for how to personalise AI assistance policy to them.

We make the following contributions:

(1) We argue for and demonstrate the importance of adapting AI assistance to hidden qualities of people. These qualities may change for a single person depending on the task and day (unlike traits, which remain constant for a longer time period). These qualities may also be hidden, meaning we must estimate them from interactions between the human and the AI assistance. This may require designing specific interactions.
(2) We focus on one specific hidden quality: overreliance rate. We show that, by including probe questions, we can quickly and accurately learn a person's hidden overreliance quality. We adapt to overreliance quality, and find that we can improve human-AI team accuracy, in particular for the overrelier group.
(3) We find that the group of people that overrely more on AI assistance put in significantly less effort, feel less pressure, and have lower perceived choice. We show that this group of people benefit from different AI assistance than the group who do not overrely on AI assistance.
(4) We find that a participant's strategy for how to use AI changes depending on what AI assistance they see in the first part of the study. This suggests that we should explicitly model a participant's strategy for using AI when adapting AI assistance to people.

2 Related Work

AI-assisted human decision making. In AI-assisted human decision making, we aim to achieve human-AI complementarity, where the human-AI team has higher accuracy than either the human or AI alone. Achieving complementary performance is usually framed as achieving appropriate reliance on the AI [42, 47]: humans should not overrely on the AI input when the AI input is wrong, and should incorporate the AI input when it is correct. However, using a fixed type of AI assistance for all people and tasks has been found to lead to worse accuracy than AI-only accuracy in many settings [3, 6, 10, 19, 36, 51], often because humans are overrelying on the AI input [7, 10, 24, 27].

Adaptive (or personalised) AI assistance, which changes depending on the person and properties of the task, has shown promise to achieve complementarity, at least for some participants [5, 8, 30, 32, 34]. Adaptive AI assistance can involve inferring some hidden preference or quality of participants. Bhatt et al. [5] learn the user's preference for which model should provide a recommendation, and personalise to this using contextual bandits. Ma et al. [30] estimate the user's capability or skill on the task, and adapt AI assistance

accordingly. Buçinca et al. [8] personalise to factors such as people's Need for Cognition trait or AI uncertainty to optimise for both accuracy and human learning. Our work focusses on inferring and adapting to a user's hidden overreliance quality.

Prior research has argued that people may overrely on AI because they superficially process AI explanations and recommendations [7, 18, 31]. Some works have used AI assistances that force humans to engage more with the AI input, and we can adaptively show these in order to achieve more appropriate reliance when needed [8, 46], although this may depend on the form of AI assistance and explanation [11]. Swaroop et al. [46] hypothesised that there is potential for using adaptive interventions after splitting people by overreliance quality but they did not test this hypothesis. We build on this work by (i) quickly inferring overreliance quality by introducing probe questions, and (ii) explicitly showing the benefit of adapting to such a hidden quality in a user study.

Using probe questions. We use probe questions as a way to estimate a participant's hidden overreliance quality. Probe questions are questions where we purposefully show an incorrect AI suggestion to the participant. Previous work has used catch trials, where an incorrect suggestion (or no suggestion at all) is shown to the user [26, 50], with the aim of ensuring the user is vigilant and alert. We use probe questions to explicitly estimate their overreliance rate, and do not use them as a method for maintaining human vigilance.

Similar to our use of probe questions, a recent concurrent work proposes using reliance drills to detect if a user is overreliant on AI [23], where their reliance drills are similar to our probe questions. Their work provides a framework for using reliance drills, providing high-level ideas for how to use results from reliance drills to ensure better human-AI team performance. Our work takes this further by also implementing probe questions in a user study, and by instantiating a method for improving performance based on overreliance quality (we adapt AI assistance depending on a participant's overreliance quality).

Reasons people have different overreliance rates. Our paper infers people's overreliance quality for the specific task, although more fine-grained versions of reliance also exist [20]. How much a person overrelies on AI input is closely related to how much a person trusts AI recommendations [7, 8] (however, in some situations an inverse relationship among these two constructs has been reported [6, 43]). Previous work has found that it is difficult to predict a person's overreliance rate on a task given personality traits such as Need for Cognition and Big-Five Personality traits [46], with work arguing that trust in AI may or may not sometimes be correlated with personality traits [38]. This could be because a person's tendency to overrely is not a stable trait, but rather depends on the specific task, the form of the AI assistance, and maybe how the person is feeling on that day (for example, how willing they are to engage with the specific task). This would mean that overreliance is not a stable trait, and indicates that we should learn overreliance quality from data in real-time, as we do in this paper.

In experiment 2 in this paper, we additionally ask participants about their Actively Open-Minded Thinking trait, which measures willingness to consider different opinions [4, 33], and therefore might be more promising than previously-used traits like Need for Cognition. We find promising results using AOT, but this will require further study and experiments.

To better understand why people might overrely on AI assistance, we ask participants what motivates them during the study. Specifically, we ask questions from the Intrinsic Motivation Inventory [40] about Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension and Perceived Choice. Given that we expect overreliers to be less motivated to engage with the task in general [7, 46], we might expect overreliers to enjoy the task less, have lower perceived competence, assign less importance to the task, feel less pressure and less perceived choice.

Importance of response time as well as accuracy. Although the focus on this paper is on the human-AI team's final accuracy, we also report response time, and explore this in more detail in Section 5. Using AI assistance to speed up decision-making time can be especially important in time-pressured settings, such as doctors in an emergency room [16, 35, 39], or in aviation [34, 41], and there can generally be an accuracy-time trade-off when choosing AI assistance [46]. Previous work has found that response time is slower on more difficult questions without necessarily increasing accuracy [1, 29], while AI assistance types that force people to engage unsurprisingly also slow down response time [46]. Our learnt personalised policy maximises decision-making accuracy, but also aims to reduce response time, and we discuss its impact on response time in our results and in Section 5.

3 Experiment 1: Probe questions help to quickly personalise

In this first experiment, we use probe questions (where we purposefully show an incorrect AI recommendation) to quickly determine a participant's hidden overreliance quality. Our first hypothesis is that using a participant's overreliance on probe questions is better than using other available information (e.g., response time, reliance rate on the AI, and personality traits) to predict the participant's overreliance quality. Directly measuring overreliance rate on probe questions (where we deliberately ensure the AI suggests a wrong answer) should be a better predictor of overreliance quality than other predictors. We classify overreliers and not-overreliers ('overreliance quality') by splitting participants into half by their overreliance rate, with the group with higher overreliance rate as the overreliers (and the group with lower overreliance rate as the not-overreliers). Note that determining a participant's true overreliance requires knowing if the AI was right or wrong on all questions, which is information we do not have outside of controlled settings, and we therefore want to accurately predict overreliance using other available information.

H1: Using probe questions distinguishes quickly between overreliers and not-overreliers, compared to only using other available information (e.g., response time, reliance on AI, personality traits).

We also hypothesise that certain probe questions are better than others at quickly determining overreliance quality. We hypothesise that easy probe questions are better than hard questions because only overreliers would rely on the AI input on easy questions; conversely, on hard questions, all people are more likely to rely on the AI input [47, 52].

H2: Easy probe questions better distinguish between overreliers and not-overreliers compared to hard questions.

We hypothesise that AI-before probe questions are better than AI-after probe questions. This is because, on AI-before questions, participants immediately choose whether or not to rely on the AI. Conversely, on AI-after, participants will already have engaged with the question and so are more likely to not rely on the AI.

H3: AI-before probe questions better distinguish between overreliers and not-overreliers compared to AI-after questions.

We also have research questions for this study. Firstly, we explore if the personalised policy in the second half leads to improved accuracy compared to baselines (maladaptive policy, and AI-before only policy), given a participant's overreliance quality. We also explore if this leads to improved response time per question. Secondly, we expect that participants who overrely more will have lower perceived choice according to the Intrinsic Motivation Inventory (IMI) questions asked post-study, and we want to explore if there are other differences in IMI answers depending on overreliance quality. Thirdly, we expect overreliers to have lower Need for Cognition trait and lower neuroticism (marginal correlations were previously found [46]). Fourthly, we want to explore if one type of probe question (e.g., easy questions with AI-before assistance) is better than the rest. Fifthly, we want to see if adding additional information (such as time taken on all questions) improves upon using our best probe question when predicting overreliance quality.

RQ1: Does personalising in second half lead to improved accuracy compared to not personalising? How is response time affected?

RQ2: Do overreliers have lower perceived choice? What about Interest/Enjoyment, Perceived Competence, Effort/Importance and Pressure/Tension?

RQ3: Do overreliers have lower NFC and lower neuroticism? What about the other Big-Five personality traits?

RQ4: Is one type of probe question better than the others?

RQ5: Does combining the best-performing type of probe question with more information (eg time taken on all questions, reliance on all questions, or personality traits) improve ability to distinguish between overreliers and not-overreliers?

3.1 Task description

We used a task setup very similar to previous work [46], where users were asked to prescribe medicines to sick fictional aliens. Participants were shown a series of sick aliens for 20 minutes, and asked to prescribe a single medicine to each alien. Each alien corresponded to a logic puzzle that the participant had to solve. This decision task is therefore accessible to laypeople while carrying real-world resemblance: we aimed to motivate participants to obtain high accuracy while getting through as many sick patients as possible.

An example task is shown in Figure 1. Participants must use the observed symptoms and the alien's 'treatment plan' (decision set rules unique to each alien) to prescribe a single medicine. We use decision sets as they are relatively easy to parse [28]. AI assistance is shown in a red box, and is provided before (AI-before) or after (AI-after) the participant's initial prescription.

There were two levels of difficulty of questions: easy and hard. Easy questions required less cognitive effort for a human to find the best medicine, while appearing superficially similar to hard questions (e.g., similar length of lines and number of lines). Figure 1 is an example of an easy question, while Figure 3 is a hard question. Additionally, we allowed for two possible correct medicines for aliens: a better medicine that treated more of the alien's observed symptoms, and a suboptimal medicine. Including a suboptimal medicine allowed us to analyse how participants use the AI input: if they simply verified that the AI recommendation was correct, then they would overrely on the suboptimal medicine (without finding the better medicine). An alternative strategy is to ignore the AI input to find the better medicine (and potentially later confirm if their better medicine treated more observed symptoms than the recommended suboptimal medicine).

We note that the task is in a fictional setting, but this allows us to precisely manipulate various aspects, such as task difficulty and AI assistance correctness. We also know the ground truth answers, allowing us to personalise AI assistance to different participants using ground-truth overreliance quality. Overall, this allows us to understand how personalisation is impacted in settings where users must complete many tasks in a limited time period.

3.2 Conditions

All our participants have see probe questions, which we use to answer Hypothesis H1. For Hypotheses H2 and H3, we designed our study with four between-subject conditions, randomly assigning participants to have one type of probe question (2 difficulties of questions x 2 AI assistance types):

(1) *Easy+Before*: Probe questions are of lower difficulty, and AI-before assistance is provided on the probe questions, meaning an AI recommendation and explanation is provided to the participant before they make a decision.
(2) *Easy+After*: Probe questions are of lower difficulty, and AI-after assistance is provided on the probe questions, meaning an AI recommendation and explanation is only provided to the participant after they make an initial decision.
(3) *Hard+Before*: Probe questions are of higher difficulty, and AI-before assistance is provided on the probe questions.
(4) *Hard+After*: Probe questions are of higher difficulty, and AI-after assistance is provided on the probe questions.

Probe questions were shown to participants every fourth question (starting on the third question): we do not want to show too many probe questions as that may take too much time away from answering non-probe questions. Probe questions were shown to participants in the first half of the study only (first 10 minutes). This is because we are interested in inferring participant quality as quickly as possible, and within the first half (and ideally much quicker).

For the second half of the study, we randomly assigned participants to one of three conditions, in order to see if adapting to overreliance rate helps (Research Question 1):

(1) *Personalised policy*: we first calculate if the participant is an overrelier or not (using true overreliance rate on all questions in the first half of the study [46]), and show the participant

Figure 1: Participants see a series of aliens, like this one, and must prescribe a single medicine to each one. Each alien is set up as a logic puzzle, with a set of inputs (the 'observed symptoms') and a set of rules (the 'treatment plan') that leads to a medicine. Participant must choose a medicine that uses only observed symptoms (and potential green 'intermediate symptoms'), and not other unobserved symptoms. We show AI assistance in a red box, like in this example, showing both a recommended medicine and an explanation (an intermediate symptom). In this example, the AI recommendation is the best possible, and all other medicines are incorrect. There are two timers counting down at the top of the screen: one says how much time remains in the 20-minute shift, and another counts down from a suggested time for each question (1 minute).

the policy personalised to their overreliance quality. The policy is described later in this subsection.

(2) *Maladaptive policy*: after calculating if the participant is an overrelier or not (based on the first half of the study), we use the policy personalised to the other group, hence making this maladapted.
(3) *AI-before policy*: we show participants only AI-before assistance on all questions, as is common in decision-support systems currently.

Personalised policy. We now describe how we chose the personalised policy for overreliers and not-overreliers. We assume we have access to the following states: (i) whether someone is an overrelier or not, (ii) the AI's uncertainty about its answer (we assume the AI is certain about its answer when it suggests a right or suboptimal answer, and uncertain when it suggests a wrong answer), (iii) the question difficulty (easy or hard). Each of these is a binary variable, meaning there are eight states in total, and we learn which AI assistance type (AI-before, AI-after, No-AI) is best for each of these, prioritising final decision-making accuracy.

We use data from a study in Swaroop et al. [46] to find which AI assistance type is best for each of the eight states, focussing on their data from the 'Mixed' setting in Experiment 2, as that resembles our experimental setup. For each of the eight states, we choose the AI assistance type that gives significantly higher accuracy. If two assistance types have similar accuracy, then we choose the assistance type that is quicker (has shorter response time). We also run Off-Policy Evaluation [45], a technique in reinforcement learning, to find the better AI assistance type when there is no significantly better assistance type for a state. We summarise our off-policy evaluation method in Appendix A.

The personalised policy is summarised in Table 1. For not-overreliers, our policy is similar to the AI-before only policy, except

Overrelier quality	AI uncertainty	Question difficulty	AI assistance type
Not-overrelier	Low	Easy	AI-after
	Low	Hard	AI-before
	High	Easy	AI-before
	High	Hard	AI-before
Overrelier	Low	Easy	AI-before
	Low	Hard	AI-before
	High	Easy	No-AI
	High	Hard	No-AI

Table 1: The personalised policy. We have three state variables and use the best AI assistance type in each state. The not-overrelier policy has mostly AI-before assistance, except when the AI is certain and the question is easy, where it has AI-after assistance. The overrelier policy shows AI-before when the AI is certain, and No-AI when the AI is uncertain.

in one state: when the AI is certain and the question difficulty is easy, we show AI-after assistance. For overreliers, our learnt policy shows AI-before when the AI is certain, and No-AI when the AI is uncertain (this forces participants to not rely on the AI assistance when the AI is uncertain).

3.3 Procedure

We conducted our study online on Prolific, a crowdsourcing platform. Participants first saw a consent form they had to accept, and then had to answer three pages of survey questions. The first page asked demographic questions, the second page asked two questions about time pressure [9] and four questions about the Need for Cognition trait (the same subset used by Gajos and Chauncey [17] from Cacioppo et al. [13]), and the third page asked the BFI-10 questions [37] to estimate participants' Big-Five Personality traits. After these survey questions, participants had to read instructions about the task, and then correctly answer three practice questions (for which they had two attempts). During the practice questions, participants were provided feedback on how to improve their answers, but no feedback was given during the main study.

After successfully completing the practice questions, participants were presented with a screen telling them to prepare for the main 20 minute study, which then began. They had to answer as many questions as possible in these 20 minutes. Participants saw an equal number of easy and hard questions, in a random order. There were two timers at the top of the screen: one global timer that counted down from 20 minutes, and a local timer giving participants 1 minute per question (the local timer turned red after reaching 0 seconds, and the timer then counted into negative numbers). For AI-after assistance, we provided 20 seconds after participants provided an initial response. There was no requirement for participants to answer each question within the 1 minute timer, but the presence of the timer increased time pressure. The global timer ensured participants felt pressured at the end of the study, while the local timer ensured there was some time pressure spread evenly throughout the study, rather than feeling relaxed during the first part.

After completing the main part of the study, participants were shown another two pages of survey questions. The first page asked 17 questions from the Intrinsic Motivation Inventory [40], about Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, and Perceived Choice (questions we used are in Appendix B). The second page asked how helpful participants found the AI (5-point scale), and asked open-ended questions on what their strategy for approaching the task was, if their strategy changed when there was an AI input, and for any other feedback.

We collected data from 207 participants on Prolific, filtering for English speakers from the US. 52 people failed the practice questions. We removed 8 people for answering questions too quickly or slowly, using the same criteria as in previous work [46]: they spent less than three seconds on at least three questions, or spent more than twice as long on one question than other questions (they got distracted for one question). This left 147 participants: to determine the required sample size, a power analysis was conducted (for Hypothesis H1), showing that 133 participants are needed to capture a medium effect size (Cohen's $\omega = 0.3$) with 80% power at a 0.05 significance level (four degrees of freedom). Each participant was paid USD$7 (median time was 37 minutes, for an estimated $11.35/hr). Failing practice questions caused the study to immediately end, and these participants were paid $2. We paid the top-performing participants a bonus $3 to motivate better performance.

Our results are based on the remaining 147 participants. Participants had a mean age of 35 years (standard deviation of 12 years). 58 participants self-identified as male, 86 as female, 2 as non-binary, and 1 preferred not to say their gender. 43 participants reported high school as their highest level of education, 63 reported a Bachelor's degree, 31 Master's (or beyond), and 10 answered 'other'.

Both experiments in our paper were approved by the Internal Review Board at Harvard University, protocol number IRB15-2076.

3.4 Design and analysis

We measure overreliance rate, looking to see if a participant's overreliance quality (determined by true overreliance rate in the first half of the study) can be predicted by overreliance rate on just probe questions.

(1) *Overreliance rate*: the proportion of times participants gave the same answer as an AI recommendation, conditioned on the AI recommendation not being optimal [7, 46, 47].

We note that our definition of overreliance rate here does not take into account the chance that a participant would give the same

answer as the incorrect AI recommendation. We do not expect this to affect our analysis because (i) this should lead to a fixed offset in overreliance rate across participants, while we are only interested in comparing overreliance between participants, and (ii) there are many potential answers for each question, and it is not likely that participants would give the same incorrect answer as the (randomly-chosen incorrect) AI recommendation.

We also see if response time can predict overreliance quality.

(2) *Response time*: we calculate how long a participant takes to answer questions on average (time taken from the moment the question is presented to the moment participant submits their answer).

In order to see if personalised policy improves accuracy (Research Question 1), we also report decision accuracy.

(3) *Accuracy*: For every question, participants can give the best answer, a suboptimal answer, or a wrong answer. The best answer corresponds to a score of 1, a suboptimal answer to a score of 0.5, and a wrong answer to a score of 0. For every participant, we calculate average accuracy across questions.

We ensured the AI accuracy was close to 0.70, in order to keep results comparable to previous work [46]. On every question, the AI had a 60% chance of recommending the best answer, a 20% chance of recommending a suboptimal answer, and a 20% of recommending a wrong answer. As participants answered different numbers of questions in their 20-minute study (depending on how quickly they answered questions on average), the overall AI accuracy varied slightly between participants. We therefore report accuracy relative to AI.

(4) *Accuracy relative to AI*: We calculate a participant's average accuracy, and subtract the average accuracy of the AI assistance they were shown. This leaves the participant's accuracy relative to the AI accuracy.

To determine if probe questions are important for determining a user's overreliance quality (hypothesis H1 and research question 5), we used a logistic regression model with true overreliance quality as the dependent variable, and available variables (average response time, reliance on AI input, personality traits (NFC and BFI traits)) and 'overreliance on probe questions' as independent variables. We ran a χ^2 test to see the importance of including overreliance on probe questions as an input variable.

To see if certain types of probe questions (hypotheses H2 and H3, and research question 4) are better than others, we first train separate logistic regression models on the different probe question types (matching how we will use probe questions in experiment 2 later, as we will just pick one type of probe question and its corresponding model). We then use a logistic regression model with response variable as 'did we successfully predict the user's overreliance quality', and dependent variable as 'probe question type'. We then report the significance of the effect (if the coefficient is not zero).

To compare performance of different policies in the second half of the study (research question 1), we split the analysis by overreliance group; we then used analysis of variance, and compared the personalised policy's performance to the two baselines (maladaptive and AI-before policies), using the Holm-Bonferroni correction method for multiple hypothesis testing [22].

To compare post-study questionnaire responses between the two groups, we used Welch's t-test (correcting for the five IMI attributes using the Holm-Bonferroni method). To see if Need for Cognition or BFI traits (based on questions asked at the beginning of the study) predict overreliance quality, we test if Pearson's correlation coefficient is not zero.

3.5 Results

3.5.1 Probe questions significantly help to distinguish between overreliers and not-overreliers (hypothesis H1). Using overreliance on probe questions significantly improves our ability to predict whether a participant is an overrelier or not. We see a significant effect of including overreliance on probe questions on top of other variables: average response time, reliance on AI and personality traits (NFC and Openness: note that Openness was the only significant BFI trait, as seen in the results of research question 3). We see this after any number of probe questions, and crucially see this early on in the study, confirming that probe questions are important to quickly determine a user's overreliance quality: overreliance on probe questions is significant in predicting a participant's overreliance quality after one probe question ($\chi^2(1, N = 147) = 39.31, p \ll .0001$), two probe questions ($\chi^2(1, N = 147) = 9.86, p = .002$), three probe questions ($\chi^2(1, N = 147) = 5.37, p = .02$), and at the half-way point in the study (10 minutes in) regardless of how many probe questions were answered ($\chi^2(1, N = 147) = 5.55, p = .02$).

3.5.2 Easy probe questions are better than hard questions at distinguishing between overreliers and not-overreliers (hypothesis H2). We find a significant effect of probe question difficulty, with easy questions better than hard questions ($p = .015$ after two probe questions, with similar results for if after one or three probe questions).

3.5.3 AI assistance type (AI-before or AI-after) does not significantly help distinguishing between overreliers and not-overreliers (hypothesis H3). We hypothesised that AI-before would be better than AI-after, but do not find a significant effect ($p = .45$ after two probe questions, with similar results for if after one or three probe questions). In fact, we find that AI-after is marginally better than AI-before.

3.5.4 Research question 1: personalised policy marginally helps overreliers. We see if the personalised policy leads to improved accuracy (relative to AI accuracy) compared to the two baselines (maladaptive and AI-before policies) after we show different policies (in the second half of the study). We note that this analysis is underpowered in this experiment, and is instead the focus of experiment 2 in Section 4. Here, we see promising trends in the results (summarised in Table 2). Across all participants, all policies have similar 'accuracy relative to AI' ($F(2, 138) = 0.29, ns$). For overreliers specifically, we observe a marginally significant main effect of policy on accuracy relative to AI ($F(2, 67) = 2.68, p = .08$): the personalised policy has similar accuracy relative to AI compared to the maladaptive policy ($p = 0.59$) and marginally higher accuracy relative to AI compared to the AI-before policy ($p = .07$). For not-overreliers, all policies have similar accuracy relative to AI ($F(2, 68) = 1.37, p = .26$).

Metric	Policy	All participants	Overreliers	Not-overreliers
Accuracy relative to AI	**Personalised**	0.03(0.02)	0.04(0.02)	0.03(0.04)
	Maladaptive	0.01(0.02)	0.02(0.02)	-0.00(0.04)
	AI-before	0.02(0.02)	-0.03(0.02)	0.08(0.03)
		$F(2, 138) = 0.29, ns$	$F(2, 67) = 2.68, p = .08$	$F(2, 68) = 1.37, ns$
Response time (s)	**Personalised**	34.5(2.0)	27.3(2.4)	41.9(2.3)
	Maladaptive	39.0(2.4)	35.8(2.9)	42.1(3.8)
	AI-before	36.1(2.2)	25.8(1.9)	46.1(2.8)
		$F(2, 138) = 0.98, ns$	$F(2, 67) = 4.50, p = .01$	$F(2, 68) = 0.64, ns$

Table 2: Mean (standard error in parentheses) of accuracy relative to AI and response time for our three policies. We look at performance over all participants, and also split into overreliers and not-overreliers. We see promising signals for the next experiment (experiment 2): the personalised policy has marginally higher accuracy relative to AI and quicker response time than the other policies. This is especially the case for overreliers. The statistics depict the main effect of policy; see text for details on statistical analysis.

We also look at the effect on response time per question. Across all participants, all policies have similar response time ($F(2, 138) = 0.98, ns$). For overreliers specifically, there is an significant effect of policy on response time ($F(2, 67) = 4.50, p = .01$), and the personalised policy leads to marginally quicker response time compared to the maladaptive policy ($p = .07$) and similar response time compared to the AI-before policy ($p = .62$). For not-overreliers, all policies have similar response time ($F(2, 68) = 0.64, ns$).

3.5.5 Research question 2: overreliers put in less effort, feel less pressure, and feel like they have less perceived choice. When looking at the IMI responses (questions asked at the end of the study), we find significant differences between the overrelier and not-overrelier groups. Overreliers answered that they put in less effort/importance ($t(145) = 2.63, p = .028$), feel less pressure/tension ($t(145) = 3.93, p < .001$), and feel like they have less perceived choice ($t(145) = 3.96, p < .001$). There are non-significant effects regarding lower interest/enjoyment ($t(145) = 1.29, ns$) and higher perceived competence ($t(145) = 1.63, ns$).

3.5.6 Research question 3: overreliers have lower Need for Cognition trait, and lower Openness trait. We see if there are correlations between traits (based on questions asked before the main study on Need for Cognition and BFI personality traits) and overreliance quality, to see if we can use these traits to help predict if a participant's overreliance quality. We find overreliers have lower Need for Cognition ($r(145) = -0.41, p = .038$) and lower Openness trait ($r(145) = -0.47, p = .020$), with the other BFI personality traits having no significant correlation with overreliance quality.

3.5.7 Research question 4: easy probe questions with AI-after assistance is marginally best. We see that the best type of probe question is Easy+After, finding it is marginally better than the other three types for distinguishing between overreliers and not-overreliers. After two probe questions, Easy+After is non-significantly better than Easy+Before and Hard+After, and significantly better than Hard+Before ($p = .03$), with similar results for after one probe question. After three probe questions, Easy+Before becomes non-significantly better than Easy+After, but we do not use this as it takes a long time to answer three probe questions (8 minutes on average), and we want to adapt our AI assistance quicker than this (two probe questions are answered after 6 minutes on average).

3.5.8 Research question 5: using probe questions with other factors (such as response time, reliance, and personality traits) does not lead to significant improvement. Finally, we see if adding information other than overreliance on probe questions can help distinguish between overreliers and not-overreliers. We consider adding response time, reliance on AI, and personality traits (specifically, NFC and Openness, as these were found to be significantly correlated with a participant's overreliance quality), and find that none of these help: overreliance on probe questions captures all the signal available from these. We train a logistic regression model to predict overreliance quality using overreliance from two probe questions, and add in our other variables. We find that our accuracy at predicting overreliance quality is always 92%, with or without any (or all) of the additional variables, indicating that overreliance on our probe questions is the only variable required.

4 Experiment 2: Personalising quickly improves performance

In the previous experiment, we saw that using probe questions can quickly distinguish between overreliers and not-overreliers. In this experiment, we use this approach to improve performance earlier than half-way through the study. We use the most predictive type of probe question from the previous study: easy questions and AI-after. We show two probe questions before classifying people as an overrelier or not-overrelier (their overreliance quality), which in experiment 1 we were able to do with an accuracy of 92%, and then show different policies to participants. On average across participants, this is after 6 minutes (instead of always after 10 minutes as in the previous study).

We make the following hypotheses informed by the first experiment's results of using personalised policies,

H4: Personalising to overreliers and not-overreliers improves human-AI team accuracy.

H4.1: Personalised policy is better than all other policies (AI-before only, random, maladaptive) for overreliers.

H4.2: Personalised policy is at least as good as all other policies (AI-before only, random, maladaptive) for not-overreliers.

Similar to the previous study, we also pose some research questions. Firstly, we explore if the personalised policy has improved average response time (as opposed to just improving accuracy). Secondly, we again look at the Intrinsic Motivation Inventory questions, and see if, like in the first study, overreliers report they put in less effort, feel less pressure, and feel like they have less perceived choice. Thirdly, we see if overreliers have lower Need For Cognition trait and lower Openness, and also add another trait that we think might correlate with overreliance quality: Actively Open-minded Thinking (AOT), which measures willingness to consider different opinions [4, 33], asking 7 questions [21].

RQ6: Does personalising in second half lead to reduced response time compared to not personalising?

RQ7: Do overreliers report they put in less effort, feel less pressure, and feel like they have less perceived choice? What about the other IMI questions?

RQ8: Do overreliers have lower NFC and lower Openness traits? What about the other Big-5 personality traits, and Actively Open-minded Thinking?

4.1 Task description and conditions

We used the same task setup as in experiment 1 and as described in Section 3.1. We showed participants a random AI assistance policy until the second probe question (both probe questions were easy questions with AI-after assistance), and then assigned participants to one of four conditions:

(1) *Personalised policy*: we infer if the participant is an overrelier or not (using performance on the first two probe questions), and show the participant the policy personalised to their overreliance quality. The policy is described in Section 3.2.
(2) *Maladaptive policy*: after inferring if the participant is an overrelier or not, we use the policy personalised to the other group, hence making this maladapted.
(3) *AI-before policy*: we show participants only AI-before assistance on all questions, as is common in decision-support systems currently.
(4) *Random policy*: for each question, we randomly choose AI-before, AI-after or no-AI assistance.

4.2 Procedure

The procedure is similar to that of experiment 1. We added questions about a participant's Actively Open-minded Thinking along with the questions about Need For Cognition (in the second page of questions, before the main study).

We collected data from 652 participants on Prolific, filtering for English speakers from the US. We chose this number by considering a small effect size ($f = 0.12$ at 80% power), which indicated we would need 548 participants in total. 112 people failed the practice questions. We removed 14 people for answering questions too quickly or slowly, using the same criteria as in the first experiment. Each participant was paid USD$7 (median time was 37 minutes, for an estimated $11.35/hr). Failing practice questions caused the study to immediately end, and these participants were paid $2. We paid the top-performing participants a bonus $3 to motivate better performance.

Our results are based on the remaining 526 participants. Participants had a mean age of 35 years (standard deviation of 12 years). 212 participants self-identified as male, 293 as female, 18 as non-binary, and 3 preferred not to say their gender. 166 participants reported high school as their highest level of education, 232 reported a Bachelor's degree, 94 Master's (or beyond), and 34 answered 'other'.

4.3 Design and analysis

We use the same metrics as in experiment 1 (overreliance rate, response time, accuracy and accuracy relative to AI).

We used similar statistical analyses as in experiment 1. To compare performance of policy after we adapt the policy to the participant, we treated each overrelier group differently, using a linear model; we then used analysis of variance, and compared the personalised policy's performance to the three baselines (maladaptive, AI-before and random policies), using the Holm-Bonferroni correction method for multiple hypothesis testing. To compare post-study questionnaire responses between the two groups, we did Welch's t-test (and corrected using the Holm-Bonferroni method). To see if Need for Cognition, AOT or BFI traits predict overreliance quality, we used a logistic regression model and ran a χ^2 test.

4.4 Results

4.4.1 Hypothesis 4: overall, personalised policy does not significantly improve upon the baseline policies. Results are in Table 3. Across all participants, the personalised policy has similar accuracy relative to AI ($F(3, 494) = 2.02, p = .11$) as baseline policies. We next split participants into overreliers and not-overreliers to see potential benefits of the personalised policy for these two groups separately.

4.4.2 Hypothesis 4.1: for overreliers, the personalised policy improves accuracy relative to AI compared to baseline policies. Results are in Table 3. For the overrelier group, the personalised policy has better accuracy relative to AI ($F(3, 222) = 12.35, p =< .001$) compared to baseline policies, with increased performance over the maladaptive policy ($p = .006$), the AI-before policy ($p = .006$) and the random policy ($p = .001$). This shows that our personalised policy benefits the overrelier group.

4.4.3 Hypothesis 4.2: for not-overreliers, the personalised policy has similar performance to the baseline policies. Results are in Table 3. For the not-overrelier group, all policies have similar accuracy relative to AI ($F(3, 268) = 0.99, ns$), indicating that no specific policy leads to a significant improvement in performance for this group of people.

4.4.4 Research question 6: Personalised policy does not speed up decision-making, and has the same response time as other policies. Across all participants, the personalised policy has similar response time to baseline policies ($F(3, 494) = 4.26, p = .006$), with all pairwise comparisons between the personalised policy and baseline policies being non-significant. For overreliers, the personalised policy does have different response time ($F(3, 222) = 3.85, p = .01$), and is slightly quicker than the maladaptive policy ($p = .19$), very slightly slower than the AI-before policy ($p = ns$), and significantly

Metric	Policy	All participants	Overreliers	Not-overreliers
Accuracy relative to AI	**Personalised (P)**	0.04(0.01)	0.03(0.01)	0.05(0.02)
	Maladaptive (M)	0.03(0.01)	-0.03(0.01)	0.08(0.02)
	AI-before (B)	0.04(0.01)	-0.03(0.01)	0.09(0.02)
	Random (R)	0.00(0.02)	-0.10(0.02)	0.09(0.02)
		$F(3, 494) = 2.02, p = .11$	$F(3, 222) = 12.35, p =< .001$	$F(3, 268) = 0.99, ns$
		—	P > {M, B, R}	—
Response time (s)	**Personalised (P)**	42.2(1.6)	32.5(1.8)	50.6(2.2)
	Maladaptive (M)	38.7(1.4)	37.5(2.3)	39.8(1.6)
	AI-before (B)	38.1(1.4)	31.0(2.1)	43.5(1.4)
	Random (R)	44.1(1.2)	39.1(1.7)	48.6(1.6)
		$F(3, 494) = 4.26, p = .006$	$F(3, 222) = 3.85, p = .01$	$F(3, 268) = 8.16, p < .001$
		n.s.	P < {R}	P > {M, B}

Table 3: Mean (standard error in parentheses) of accuracy relative to AI and response time for our four policies. We look at performance over all participants, and also split into overreliers and not-overreliers. We see that, for overreliers, the personalised policy significantly improves accuracy relative to AI (hypothesis H4.1), but over all participants this is not significant (hypothesis H4). As a research question (RQ6), we also look at response time, finding that the personalised policy does not reduce response time (and in fact reduces response time for not-overreliers). See text for details on statistical analysis.

quicker than the random policy ($p = .02$). For not-overreliers, the personalised policy is slower than baseline policies ($F(3, 268) = 8.16, p < .001$), and is significantly slower than the maladaptive policy ($p = .0002$) and the AI-before policy ($p = .02$), and is similar to the random policy ($p = ns$). This result for not-overreliers is surprising as we expected all policies' performance to be similar, as in the results from experiment 1. We discuss possible reasons for this in Section 5.

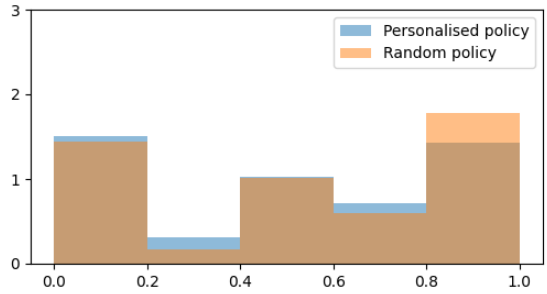
4.4.5 Research question 7: like in experiment 1, overreliers put in less effort and feel like they have less perceived choice. We find similar results from the IMI responses as in experiment 1. Overreliers answered that they put in less effort/importance ($t(524) = 2.94, p = .014$), feel like they have less perceived choice ($t(524) = 6.55, p < .001$), marginally feel less pressure/tension ($t(524) = 2.00, p = .14$), with non-significant effects regarding lower interest/enjoyment ($t(524) = 1.21, ns$) and higher perceived competence ($t(524) = 1.10, ns$). Compared to experiment 1, the only change is that there is a marginal effect of overreliers feeling less pressure/tension (this was a significant effect in experiment 1). However, we note that in experiment 1, we used a different definition of overreliance quality: here, we are only using two probe questions to determine overreliance quality (which was 92% accurate in experiment 1), and this slight difference may explain small changes in effects.

4.4.6 Research question 8: unlike experiment 1, overreliers do not have significantly lower NFC or Openness; they have higher Agreeableness, Neuroticism and lower Actively Open-minded Thinking. We see if there are correlations between traits (based on questions asked before the main study) and overreliance quality, to see if we can use these traits to help predict a participant's overreliance quality. We find overreliers have marginally lower Need for Cognition ($r(524) = -0.12, p = .13$), higher Agreeableness trait ($r(524) = 0.21, p = .022$), and lower Neuroticism ($r(524) = -0.25, p < .001$), with the other BFI personality traits having no significant correlation with overreliance quality. This is slightly different to the results from experiment 1, highlighting that it is difficult to use these traits to predict a person's overreliance quality. In this experiment, we also estimated a participant's Actively Open-minded Thinking (AOT) trait. We find that overreliers have significantly lower AOT ($r(524) = -0.51, p < .001$), indicating that we should include AOT in future experiments.

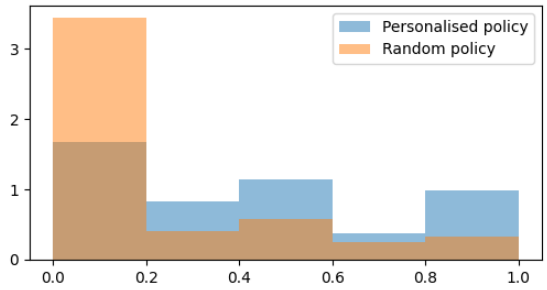
5 Exploratory Analysis: different AI assistance policies lead to using AI in different ways

In this section, we look in more detail at the per-question response time of participants. We explore why, for not-overreliers, the personalised policy in experiment 2 made participants slower than the maladaptive policy, while in experiment 1, they had similar response times. We find that, by adapting the policy to overreliance quality earlier in the study in experiment 2, we affect how participants use the AI input. This in turn changes what policies are best for participants. We see this effect on not-overreliers because they engage with the task more, therefore changing how they use AI assistance depending on the policy shown. Overreliers engage less with the task, and the best AI assistance policy for them is not affected as much by which AI assistance policies they see early during the study.

Using overreliance rates to determine participants' strategy. In this section, we use overreliance rate on AI to make statements about the strategies participants learn for using the AI assistance. We look at overreliance rate on the AI suggestion when the AI is suboptimal, and separately the overreliance rate when the AI is wrong (in Section 3.4 we defined overreliance rate as the combination of these two). We always look at these overreliance rates over the second half of the study only, so that we can compare overreliance rates between the two experiments. By looking at the overreliance rates on suboptimal and wrong AI suggestions separately, we can draw conclusions of how participants use the AI assistance. If participants have high overreliance when the AI is suboptimal, but low when the AI is wrong, this suggests that they are simply verifying the



(a) Overreliance on suboptimal AI.



(b) Overreliance on wrong AI.

Figure 2: Histogram plots (normalised density) of overreliance rate on suboptimal AI and wrong AI for the personalised policy and random policy in experiment 2 for the not-overrelier group (all policies in experiment 1 have similar plots to the random policy in experiment 2 in this figure). We see that the overreliance rate on suboptimal AI is similar, however, there is more of a spread of overreliance on wrong AI for the personalised policy. This indicates that participants with the random policy learn to verify the AI input, while participants with the personalised policy learn a mix of strategies.

AI suggestion: after confirming that the AI's suggestion is right (like a suboptimal answer), they do not check if a better answer exists among the choices. Alternatively, if participants have low overreliance when the AI is suboptimal and low overreliance when the AI is wrong, this suggests they are ignoring the AI assistance entirely.

Experiment 1 strategies. In experiment 1, participants see a random policy in the first half of the study. When we look at participants' overreliance rates in the second half, we see that not-overreliers learned to verify the AI input. All policies have overreliance rates on AI suboptimal in the range 0.57-0.60 (the overall AI-suboptimal overreliance rate is 0.58 ± 0.05), while all policies have overreliance rates on AI wrong of 0.22 (the overall AI-wrong overreliance rate is 0.22 ± 0.05). This indicates that participants verify the AI input, regardless of policy seen in the second half of the study. This verification strategy leads to good performance using the personalised policy, and marginally worse performance using the maladaptive policy, as reported in Section 3: the personalised policy shows AI input even when the AI is uncertain or wrong, and verifying an incorrect answer is quick (and can even save time as it can require partially completing the logic puzzle already).

Experiment 2 strategies. On the other hand, in experiment 2, participants see a random policy only until the second probe question, after which the policy immediately changes (as opposed to changing only in the second half of the study). This does not appear to be long enough for not-overrelier participants to learn to verify the AI input. Instead, they learn different strategies depending on the policy they are changed to. When shown a random policy, their strategy is similar to experiment 1 (as we would expect). However, when shown the personalised policy or the AI-before policy, participants on average rely on the AI input more. In fact, there is much more of a spread of strategies: some participants learn to verify the AI input, but others either completely ignore the AI, or overrely on it.

We can see this by looking at overreliance rates: for a random policy, participants have similar overreliance rates as in experiment 1 (AI-suboptimal overreliance rate 0.55 ± 0.05, and AI-wrong overreliance rate 0.18 ± 0.04). For the personalised policy, there is higher AI-wrong overreliance rate (0.41 ± 0.05 for the personalised policy), and this reduces decision-making accuracy. We can see the spread of strategies by looking at the histogram of overreliance rates for the personalised policy compared with the random policy in experiment 2 (Figure 2). Overreliance on suboptimal AI suggestions is similar, but overreliance on wrong AI is much more spread out for the personalised policy: this indicates that, with the random policy, participants learn to verify the AI input (high overreliance on suboptimal AI, low on wrong AI); but with the personalised policy, participants also overrely on the AI (high overreliance on both suboptimal AI and wrong AI) and ignore the AI (low overreliance on both suboptimal AI and wrong AI). On average across participants, this mix of strategies also leads to the increased per-question response time: the participants that ignore the AI input are much slower than all other participants.

Additionally, not-overreliers learn to use the maladaptive policy very well: they overrely on the AI input (overreliance rate 0.61 ± 0.04), but this does not affect accuracy significantly, as this policy does not show an AI input when the AI is uncertain (when the AI recommendation is wrong). We can see this, for example, through the higher reliance on AI when the AI input is suboptimal (0.61 ± 0.04 for the maladaptive policy, and 0.51 ± 0.05 for the personalised policy). This increased reliance allows participants to significantly speed up the decision-making, leading to improved performance with this maladaptive policy compared to other policies.

Participants prefer the policy that they performed better with. We also see this different use of policy in participants' post-study questionnaire, specifically how helpful they found the AI assistance. In experiment 1, not-overreliers marginally found the personalised policy (helpfulness 0.24 ± 0.18) to be more helpful than the maladaptive policy (helpfulness -0.15 ± 0.22) and the AI-before policy (helpfulness 0.05 ± 0.26). This is because they were able to use the personalised policy to verify the AI input, speeding up decision-making. In experiment 2, not-overreliers found the maladaptive

policy (helpfulness 0.73 ± 0.14) significantly more helpful than all other policies: the personalised policy (helpfulness -0.03 ± 0.12), AI-before policy (helpfulness 0.24 ± 0.11), and random policy (helpfulness 0.07 ± 0.13). This is because they did not learn to verify the AI input, and so did not find the personalised policy helpful; instead, they learnt how to use the maladaptive policy.

Overall, the analysis in this section suggests that people who engage with the task (the not-overrelier group) actively model and use the AI input in different ways. The type of assistance they are shown early on affects how they use AI assistance throughout the rest of the study, and they do not update how they use AI assistance later in the study. It is therefore important to actively model how participants use or view the AI input: for example, are they learning to verify the AI input, or are they ignoring it? Alternatively, we could explicitly include the types of AI assistance previously shown in our own model of the human-AI team. This time-dependency also suggests that we could use a full reinforcement learning model, instead of stationary bandit policies.

6 Discussion

In this paper, we look at how to personalise to people's hidden overreliance quality. Previous work has looked at the benefits of personalising AI assistance to other qualities of people, using different policies for different people [5, 8, 30]. We focus on overreliance quality, which past work has found is a relevant stable quantity during a study [46]. We want to personalise to this quality as quickly as possible in order to improve human-AI team performance, and we introduce probe questions as a way to do so.

6.1 Personalising to overreliance rate improves performance

We find that personalising to our hidden quality (overreliance) can improve accuracy for the group of people that overrely on the AI assistance, while all AI assistance policies perform similarly well for the not-overrelier group. Past work has found that adapting AI assistance to people can lead to improved human-AI team accuracy [5, 8, 30, 32]. Our work adds to this literature: adapting AI assistance to different people (and to different properties of the task) improves performance of the human-AI team, helping achieve complementary performance.

We find that overreliers put in less effort / assign less importance to the task, feel less pressure, and feel like they have less perceived choice than not-overreliers in both our experiments. This shows that the two groups of people are different in how they approach the task, and it may therefore be unsurprising that we should show different AI assistances to them. Previous work has seen that different groups of people put in less effort and/or engage less, and this adds to that work [7]. The degree of engagement and effort people exert on the task may depend on (i) situational factors, such as stress or time pressure [46], (ii) semi-stable traits, such as expertise level [14], or (iii) individual differences, such as people's intrinsic motivation to think [7]. In this work, we use overreliance quality as a proxy for capturing this, and personalise policies accordingly.

Overreliance rate is one hidden quality, and other hidden qualities may be relevant in other settings. For example, we may want to estimate a user's skill on a task [8], or their preference for different forms of assistance [5]. We also note that such hidden qualities can be different for the same person depending on the task setting, making estimating the quality from real-time interactions important.

6.2 Probe questions help to quickly personalise

We focus on adapting to a *hidden* quality, meaning we have to infer the value of this quality so that we can personalise to it. This is especially difficult to do quickly. We find that using already-available metrics, such as response time per question and reliance rate on the AI, are not very good for predicting a person's hidden overreliance quality. We also find that personality traits (such as NFC and BFI traits) can correlate with a person's hidden overreliance quality, but do not capture enough signal. Instead, we introduce probe questions, where we know the correct answer, and purposefully show an incorrect AI suggestion to see if people overrely on it. Using probe questions significantly helps in inferring a person's hidden overreliance quality in our first experiment, and even two probe questions are enough to predict a person's overreliance quality with greater than 90% accuracy. In general, for different hidden qualities, it may be be necessary to use additional signals or data to personalise quickly, similar to how we introduced probe questions to quickly infer a person's overreliance quality.

We note that probe questions have been used in different ways in human-AI teams before in order to ensure sustained vigilance [26, 50], such as in catch trials or in realistic settings, such as in airport baggage screening [49]. This work shows probe questions can also be used as a mechanism to estimate a person's overreliance quality, and AI assistance can be adapted to overreliance quality in order to improve performance. We believe that when appropriately introduced in realistic settings, this mechanism can be extended to measure other human hidden qualities that may depend on wide array of factors, such as skill level.

6.3 People learn different strategies for using AI assistance

Our exploratory analysis in Section 5 suggests that people learn to use AI assistance in different ways, and that the strategy they learn depends on the AI assistance they saw early in the study. We found that, in experiment 2, the not-overrelier group (who engage more with the task) use AI assistance differently to the not-overrelier group from experiment 1. In the two experiments people were shown different AI assistance policies earlier in their study: in experiment 1, participants were shown a random policy for the entire first half of the study (10 minutes into the study), while in experiment 2, participants were shown a personalised or baseline policy after answering two probe questions (on average, 6 minutes into the study). This then appears to have impacted their performance with the personalised policy: not-overreliers were slower using the personalised policy than the maladaptive policy in experiment 2, while this was not the case for not-overreliers in experiment 1.

We find that in experiment 1, not-overrelier participants learn to verify the AI input, using the AI input to speed up decision-making. They learn this strategy after seeing a random policy in the first half of the study. This same strategy is learned for participants

in experiment 2 that saw a random policy throughout the study. However, participants in experiment 2 that saw a different policy early in the study (such as the personalised policy) seem to learn different strategies for using AI assistance. This difference in learned strategies leads to slower average response time. Our findings on changing AI assistance extend prior research in AI-assisted decision-making, which demonstrated that a change to error boundaries [2] or explanations [48] impacted people's behaviour (their objective performance on the task), as well as people's subjective experiences with the AI assistant.

Our analysis suggests that it is important to take a person's strategy for using AI assistance into account when adapting to different people. We can view overreliance quality as determining whether a participant overrelies on the AI input or not, but this quality does not explicitly determine if a participant ignores or verifies the AI to improve their answer. Consistent with prior work that has also highlighted that people may use different strategies when incorporating AI suggestions into their decision-making [44], our exploratory analysis further suggests that we could explicitly model how each person uses AI input (the different strategies they use), and potentially use this as the hidden quality we personalise to, with personalised policies depending on a person's strategy. More generally, future work can explicitly estimate how a participant models or views the AI assistant.

Our results also suggest that AI assistance policy impacts how a person uses AI input, with different policies leading to people learning different ways of using AI input (especially if the policy is different earlier on in the study, when people are still learning how to use the AI assistance). We therefore need to take into account what types of AI assistance people have seen before. This highlights the importance of a data-oriented approach (where we learn hidden qualities in real-time), as opposed to relying on pre-study surveys to estimate these qualities. We could model this in real-time using, for example, a full reinforcement learning setup, as opposed to the bandit setup we used in this work. However, this can increase the complexity of the model and be difficult to learn in noisy human settings.

6.4 Limitations

We use a setting where participants answer a series of logic puzzles, based on related work [46]. This setting may not be realistic as there is no previous knowledge that can help with this task. We also explicitly design our own AI assistance, instead of using a machine learning model. This allows us to ensure our AI assistance has similar accuracy to human-only accuracy. But we assume we have good estimates of uncertainty from the AI assistant, which may not be realistic with a trained machine learning model.

In our studies, participants did not know when a question was a probe question or not. This can be realistic in some settings (such as in some catch trial settings [50]), but may not be the case in other settings (for example, doctors will know if there is a fake patient or not).

Our task design allowed people to verify if the AI assistance was correct or not, and then use this to potentially speed up finding the best possible answer. This may not be possible in other tasks with other kinds of AI assistance (e.g., not all AI suggestions may be verifiable), meaning that we would need to consider different participant strategies for using the AI assistant.

7 Conclusion

Adapting AI assistance to people and to the task can help improve accuracy and performance. To achieve this, we need to adapt to hidden qualities of people, and this requires quickly estimating these qualities before adapting our AI assistance. In this paper, we show how we can quickly adapt to people's hidden overreliance quality.

We found that introducing probe questions, where we purposefully show an incorrect AI suggestion and see if the participant overrelies on it, helps us to quickly infer the participant's hidden overreliance quality. We then designed a personalised AI assistant policy for different people, and found that this policy especially helps the overrelier group of people (people who engage less with the task).

In our exploratory analysis, we also found that people learn different strategies for using AI assistance, and that the strategy they learn depends on what AI assistance they saw early in the study. This indicates that we can treat a participant's strategy as a hidden quality that we personalise to (and that changes over time); we could alternatively use a full reinforcement learning setup where we take into account what AI assistance was shown to participants earlier in the study.

Acknowledgments

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

[1] Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. 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* (Parkville, VIC, Australia) *(OzCHI '15)*. Association for Computing Machinery, New York, NY, USA, 352–360. https://doi.org/10.1145/2838739.2838753

[2] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 81, 16 pages. https://doi.org/10.1145/3411764.3445717

[4] Jonathan Baron. 1985. *Rationality and Intelligence.* Cambridge University Press, Cambridge.

[5] Umang Bhatt, Valerie Chen, Katherine M Collins, Parameswaran Kamalaruban, Emma Kallina, Adrian Weller, and Ameet Talwalkar. 2023. Learning Personalized Decision Support Policies.

[6] 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* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 454–464. https://doi.org/10.1145/3377325.3377498

[7] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (April 2021), 21 pages. https://doi.org/10.1145/3449287

[8] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Susan A. Murphy, and Krzysztof Z. Gajos. 2024. Towards Optimizing Human-Centric Objectives in AI-Assisted Decision-Making With Offline Reinforcement Learning.

[9] Thomas Buser, Roel van Veldhuizen, and Yang Zhong. 2022. *Time Pressure Preferences.* Technical Report. Tinbergen Institute Discussion Paper.

[10] 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, Dallas, TX, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[11] Zana Buçinca, Siddharth Swaroop, Amanda E. Paluch, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2025. Contrastive Explanations That Anticipate Human Misconceptions Can Improve Human Decision-Making Skills. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'25).* ACM, Yokohama, Japan. https://doi.org/10.1145/3706598.3713229

[12] John T. Cacioppo and Richard E. Petty. 1982. The need for cognition. *Journal of Personality and Social Psychology* 42, 1 (1982), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

[13] J T Cacioppo, R E Petty, and C F Kao. 1984. The efficient assessment of need for cognition. *Journal of personality assessment* 48, 3 (1984), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

[14] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C. Nascimento. 2023. Assertiveness-based Agent Communication for a Personalized Medicine on Medical Imaging Diagnosis. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 13, 20 pages. https://doi.org/10.1145/3544548.3580682

[15] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Springer Verlag, New York, US.

[16] Amy Franklin, Ying Liu, Zhe Li, Vickie Nguyen, Todd R. Johnson, David Robinson, Nnaemeka Okafor, Brent King, Vimla L. Patel, and Jiajie Zhang. 2011. Opportunistic decision making and complexity in emergency care. *Journal of Biomedical Informatics* 44, 3 (2011), 469–476. https://doi.org/10.1016/j.jbi.2011.04.001 Biomedical Complexity and Error.

[17] Krzysztof Z. Gajos and Krysta Chauncey. 2017. The Influence of Personality Traits and Cognitive Load on the Use of Adaptive User Interfaces. In *Proceedings of the 22Nd International Conference on Intelligent User Interfaces* (Limassol, Cyprus) *(IUI '17).* ACM, New York, NY, USA, 301–306. https://doi.org/10.1145/3025171.3025192

[18] 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* (Helsinki, Finland) *(IUI '22).* Association for Computing Machinery, New York, NY, USA, 794–806. https://doi.org/10.1145/3490099.3511138

[19] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[20] Ziyang Guo, Yifan Wu, Jason D. Hartline, and Jessica Hullman. 2024. A Decision Theoretic Framework for Measuring AI Reliance. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24).* Association for Computing Machinery, New York, NY, USA, 221–236. https://doi.org/10.1145/3630106.3658901

[21] Uriel Haran, Ilana Ritov, and Barbara A Mellers. 2013. The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision making* 8, 3 (2013), 188–201.

[22] Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[23] Rosco Hunter, Richard Moulange, Jamie Bernardi, and Merlin Stein. 2024. Monitoring Human Dependence On AI Systems With Reliance Drills.

[24] Maia Jacobs, Melanie F.Pradier, Thomas McCoy, Roy Perlis, Finale Doshi velez, and Krzysztof Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry* 11 (02 2021). https://doi.org/10.1038/s41398-021-01224-x

[25] Patricia K. Kahr, Gerrit Rooks, Chris Snijders, and Martijn C. Willemsen. 2024. The Trust Recovery Journey. The Effect of Timing of Errors on the Willingness to Follow AI Advice.. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24).* Association for Computing Machinery, New York, NY, USA, 609–622. https://doi.org/10.1145/3640543.3645167

[26] Dimitrios Kourtis, Pierre Jacob, Natalie Sebanz, Dan Sperber, and Günther Knoblich. 2020. Making sense of human interaction benefits from communicative cues. *Scientific Reports* 10, 1 (2020), 18135–.

[27] Vivian Lai and Chenhao Tan. 2019. 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* (Atlanta, GA, USA) *(FAT* '19).* Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[28] Hima Lakkaraju, Stephen Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. *KDD : proceedings. International Conference on Knowledge Discovery & Data Mining* 2016 (08 2016), 1675–1684. https://doi.org/10.1145/2939672.2939874

[29] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. 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* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 72, 13 pages. https://doi.org/10.1145/3411764.3445522

[30] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. 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* (Hamburg, Germany) *(CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 759, 19 pages. https://doi.org/10.1145/3544548.3581058

[31] Tim Miller. 2023. Explainable AI is Dead, Long Live Explainable AI! Hypothesis-Driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23).* Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[32] Gali Noti and Yiling Chen. 2023. Learning When to Advise Human Decision Makers. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23,* Edith Elkind (Ed.). International Joint Conferences on Artificial Intelligence Organization, Macao, SAR, China, 3038–3048. https://doi.org/10.24963/ijcai.2023/339 Main Track.

[33] Victor Ottati and Chadly Stern. 2023. *Divided: Open-Mindedness and Dogmatism in a Polarized World.* Oxford University Press, Oxford, UK. https://doi.org/10.1093/oso/9780197655467.001.0001

[34] Raja Parasuraman, Mustapha Mouloua, and Robert Molloy. 1996. Effects of Adaptive Task Allocation on Monitoring of Automated Systems. *Human Factors* 38, 4 (1996), 665–679. https://doi.org/10.1518/001872096778827279 arXiv:https://doi.org/10.1518/001872096778827279 PMID: 11536753.

[35] Vimla Patel, Jiajie Zhang, Nicole Yoskowitz, Robert Green, and Osman Sayan. 2008. Translational Cognition for Decision Support in Critical Care Environments: A Review. *Journal of biomedical informatics* 41 (07 2008), 413–31. https://doi.org/10.1016/j.jbi.2008.01.013

[36] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. 2018. Manipulating and measuring model interpretability.

[37] Beatrice Rammstedt and Oliver P. John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41, 1 (2007), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

[38] René Riedl. 2022. Is trust in artificial intelligence systems related to user personality? Review of empirical evidence and future research directions. *Electronic Markets* 32, 4 (2022), 2021–2051.

[39] Leonardo Rundo, Roberto Pirrone, Salvatore Vitabile, Evis Sala, and Orazio Gambino. 2020. Recent advances of HCI in decision-making tasks for optimized clinical workflows and precision medicine. *Journal of Biomedical Informatics* 108 (2020), 103479. https://doi.org/10.1016/j.jbi.2020.103479

[40] Richard Ryan and Edward Deci. 2000. Self-Determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-Being. *The American psychologist* 55 (01 2000), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

[41] Nadine B. Sarter and Beth Schroeder. 2001. Supporting Decision Making and Action Selection under Time Pressure and Uncertainty: The Case of In-Flight Icing. *Human Factors* 43, 4 (2001), 573–583. https://doi.org/10.1518/001872001775870403

[42] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making.

[43] Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. 2021. Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. https://www.alexandria.unisg.ch/handle/20.500.14171/111308

[44] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23).* Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[45] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (second ed.). The MIT Press, Cambridge, MA, US. http://incompleteideas.net/book/the-book-2nd.html

[46] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) *(IUI '24).* Association for Computing Machinery,

New York, NY, USA, 138–154. https://doi.org/10.1145/3640543.3645206
[47] 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. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (April 2023), 38 pages. https://doi.org/10.1145/3579605
[48] Xinru Wang and Ming Yin. 2023. Watch Out for Updates: Understanding the Effects of Model Explanation Updates in AI-Assisted Decision Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 758, 19 pages. https://doi.org/10.1145/3544548.3581366
[49] Jeremy M Wolfe, David N Brunelli, Joshua Rubinstein, and Todd S Horowitz. 2013. Prevalence effects in newly trained airport checkpoint screeners: Trained observers miss rare targets, too. *Journal of vision* 13, 3 (2013), 33–33.
[50] John Zerilli, Umang Bhatt, and Adrian Weller. 2022. How Transparency Modulates Trust in Artificial Intelligence. *Patterns* 3, 4 (2022), 1–10.
[51] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852
[52] Zelun Tony Zhang, Felicitas Buchner, Yuanting Liu, and Andreas Butz. 2024. You Can Only Verify When You Know the Answer: Feature-Based Explanations Reduce Overreliance on AI for Easy Decisions, but Not for Hard Ones. In *Proceedings of Mensch Und Computer 2024* (Karlsruhe, Germany) *(MuC '24)*. Association for Computing Machinery, New York, NY, USA, 156–170. https://doi.org/10.1145/3670653.3670660

A Using off-policy evaluation to learn personalised policy

In this section, we briefly describe our off-policy evaluation (OPE) technique [45] to learn the best personalised policy (the policy is summarised in Table 1).

We use available data from a previous work [46] to find which AI assistance type is best for each of the eight states, using data from their Experiment 2's 'Mixed' setting, as that resembles our experimental setup. For each of the eight states, we choose the AI assistance type that gives significantly higher accuracy. If two assistance types have similar accuracy, we choose the assistance type that is quicker (shorter response time). If there is no clear better assistance type for a specific state, we use OPE to find the better assistance type, choosing the one that leads to higher accuracy.

We set up our OPE method as follows. For each participant, we group their responses to each question by which state that question corresponds to (state is decided by whether the participant is an overrelier or not, the AI's uncertainty, and the question difficulty) and by the AI assistance type shown (AI-before, AI-after or no-AI). This grouping allows us to store each participant's accuracies for every state and AI assistance type. Then, in order to estimate how good a test policy is, we first go over each participant, sampling questions randomly. We use the test policy to decide the AI assistance shown for each question, and sample an accuracy based on the participant's stored accuracies (we earlier stored a list of accuracies for the participant for every state and AI assistance type). This then provides an accuracy per question for this participant, which we average to calculate the participant's average accuracy. We then average across participants. We repeat this method 5 times to reduce randomness due to sampling accuracies. We choose the test policy that leads to highest accuracy overall.

B Intrinsic Motivation Inventory Questions in our study

This section lists the questions we ask participants (at the end of the study) about Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, and Perceived Choice. All questions were asked using a 7-point Likert scale. Questions with reverse scoring are indicated with "(R)".

(1) **Interest/Enjoyment**
"I enjoyed finding medicines very much."
"This activity did not hold my attention at all." (R)
"While I was doing this activity, I was thinking about how much I enjoyed it."
"This activity was fun to do."
(2) **Perceived Competence**
"I am satisfied with my performance at this task."
"After working at finding medicines for a while, I felt pretty competent."
"I think I am pretty good at this activity of treating aliens."
"I think I did pretty well at this activity, compared to other participants."
(3) **Effort/Importance**
"I put a lot of effort into finding medicines."
"I didn't put much energy into finding a good medicine." (R)
"I tried very hard to treat aliens well."
(4) **Pressure/Tension**
"I felt very tense while treating aliens."
"I was very relaxed while finding medicines." (R)
"I did not feel nervous at all while doing this." (R)
(5) **Perceived Choice**
"I felt like I was strongly influenced by the AI on how to treat aliens." (R)
"I found medicines in the way I wanted to."
"I was free to choose the medicines I thought were best for each patient."

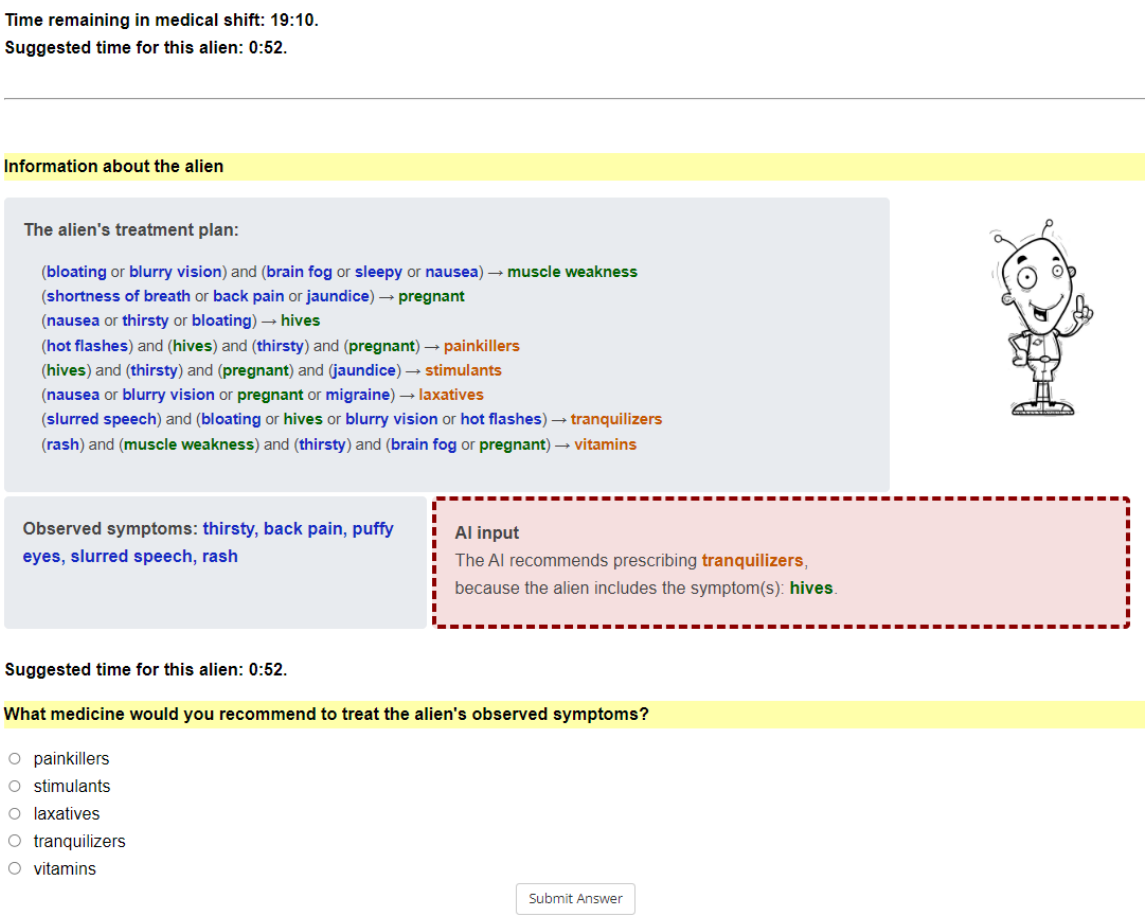


Figure 3: An example of a hard difficulty alien (compare with the easy example in Figure 1). Here, there are many medicines that treat more symptoms than the best correct medicine, meaning participants must manually check these other medicines before confirming that they are incorrect options. This takes participants more time, and leads to lower human-only accuracy on average. In this example, the recommended medicine ('tranquilizers') is the best medicine, followed by a suboptimal (but still correct) answer of 'laxatives'.