Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning

Krzysztof Z. Gajos
kgajos@eecs.harvard.edu
Harvard School of Engineering and Applied Sciences
Allston, MA, USA

Lena Mamykina
om2196@cumc.columbia.edu
Columbia University
New York, NY, USA

ABSTRACT

When people receive advice while making difficult decisions, they often make better decisions in the moment and also increase their knowledge in the process. However, such *incidental learning* can only occur when people cognitively engage with the information they receive and process this information thoughtfully. How do people process the information and advice they receive from AI, and do they engage with it deeply enough to enable learning? To answer these questions, we conducted three experiments in which individuals were asked to make nutritional decisions and received simulated AI recommendations and explanations. In the first experiment, we found that when people were presented with both a recommendation and an explanation before making their choice, they made better decisions than they did when they received no such help, but they did not learn. In the second experiment, participants first made their own choice, and only then saw a recommendation and an explanation from AI; this condition also resulted in improved decisions, but no learning. However, in our third experiment, participants were presented with just an AI explanation but no recommendation and had to arrive at their own decision. This condition led to both more accurate decisions and learning gains. We hypothesize that learning gains in this condition were due to deeper engagement with explanations needed to arrive at the decisions. This work provides some of the most direct evidence to date that it may not be sufficient to include explanations together with AI-generated recommendation to ensure that people engage carefully with the AI-provided information. This work also presents one technique that enables incidental learning and, by implication, can help people process AI recommendations and explanations more carefully.

CCS CONCEPTS

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

KEYWORDS

decision support systems, incidental learning, cognitive engagement, explainable AI, human-centered AI

IUI '22, March 22–25, 2022, Helsinki, Finland

ACM ISBN 978-1-4503-9144-3/22/03…$15.00
https://doi.org/10.1145/3490099.3511138

ACM Reference Format:
Krzysztof Z. Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces (IUI '22), March 22–25, 2022, Helsinki, Finland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3490099.3511138

1 INTRODUCTION

In many areas of human enterprise, individuals increasingly rely on Artificial Intelligence (AI) to inform their decisions and choices. There is growing evidence that people supported by such systems can, on average, make better decisions compared to the decisions they would have made on their own [11, 29, 40]. However, previous studies also showed human tendency to over-rely on AI-generated recommendations [4, 8, 34, 66], which suggests that people may be processing information provided by AI superficially rather than engaging with it deeply and critically using their own knowledge and expertise. Given continuous concerns regarding the reliability and trustworthiness of AI, human critical engagement may be a necessary component of successful human-AI interaction, particularly in domains with a high cost of errors, such as health and medicine. There is already a growing body of work showing how the interactions with AI-powered decision support systems could be redesigned—using approaches ranging from tutorials [39] to in-the-moment cognitive interventions [9, 54]—so as to encourage deeper processing of the AI-generated information.

Researchers in learning sciences use the term "cognitive engagement" to describe learners' engagement with the learning process. When people are cognitively engaged with instructional process and materials, they are more likely to benefit from instruction and are more likely to acquire new skills and knowledge. We propose that cognitive engagement may be a useful construct in conceptualizing human engagement with AI and can help to distinguish between passive engagement, when individuals simply follow AI recommendations, and deeper forms of engagement, when they critically examine these recommendations and compare them with their own knowledge and judgement. An outcome of deeper cognitive engagement would be an ability to reject information that is inconsistent with individuals' own knowledge and beliefs, and to adjust their own knowledge to incorporate new information. This type of knowledge acquisition happens not only with formal instruction, but is also common in professional settings, when individuals interact with others in order to accomplish tasks, and use these interactions to increase their own knowledge "about facts, domains, history, assumptions, strategies" related to the task [6]. This type of learning is commonly referred to as accidental learning [47, 48].

In this research, we examined the impact of different approaches to the design of human-AI interactions on incidental learning. Given previous research on human-AI interaction, we hypothesized that simply presenting a person with a decision suggestion and an explanation would provide an immediate benefit (i.e., help the person make a better decision) but would not lead to learning. However, we also hypothesized that alternative forms of the human-AI interactions—designed to elicit deeper processing of the AI-generated information—would both provide an immediate benefit and lead to incidental learning. We tested two such alternative designs. In the first one, which we refer to as the *Update* design, people first made an initial decision on their own before being shown the AI recommendation and explanation and having a chance to revise their decision. In the second one, the *AI explanation only* design, participants were shown just an AI explanation but no explicit recommendation—they had to use the information from the explanation to arrive at the optimal decision themselves.

We conducted three experiments in which participants had to make a series of nutrition-related decisions (decide which of two meals shown was a greater source of a specified macronutrient). In each experiment, we compared one human-AI interaction to two non-AI baselines (in one baseline condition, participants received simple correctness feedback on their choices; in the second, they received both correctness feedback and an explanation). In all three experiments, the AI assistance provided significantly higher immediate benefit compared to the baseline conditions where such assistance was not present. As hypothesized, the results of the first experiment (n=251), showed that simply presenting people with an AI recommendation and explanation did not result in greater learning than in the baseline design where people received no assistance and no feedback. Contrary to our expectations, the results of the second experiment (n=268), showed that the *Update* design also did not result in learning. However, the results of the third experiment (n=221 and a replication with n=300) demonstrated that the *AI explanation only* design did result in learning while also providing immediate benefit.

We hypothesize that the observed difference in learning gain was due to the degree of cognitive engagement with AI-generated information. When individuals were provided with a solution to their task (in the form of a decision recommendation), they did not need to engage deeply with the explanations and could simply proceed with action. However, when they needed to arrive at their own decisions, they needed to engage with the provided explanations and synthesize the information to arrive at the conclusions. These results have implications for future AI-powered systems for supporting human decisions: contrary to common expectations, merely providing explanations for AI recommendations may not be enough to ensure that people critically evaluate those recommendations and arrive at final decisions that appropriately combine their own knowledge and information contributed by the AI. Instead, other forms of AI support that focus on presenting useful information rather than recommendations for solutions, may elicit deeper cognitive engagement and prompt individuals to more critically and thoughtfully examine assistance from AI.

In this work, we make the following contributions:

- The results of our first experiment show that people who were offered an AI-generated decision recommendation accompanied by an informative explanation performed better on the task at hand compared to when no AI support was offered, but did not learn from the AI-provided information. Given the strong link between cognitive engagement and learning, this is some of the most direct evidence to date that it may not be sufficient to include explanations together with AI-generated recommendation to ensure that people engage carefully with the AI-provided information.
- We demonstrated that an alternative design of the human-AI interaction, one in which the AI presents just an explanation leaving the person to arrive at the decision, provides both an immediate benefit in terms of decision quality and supports incidental learning.

2 RELATED WORK

Contrary to the initial expectations [35, 36], people supported by AI-powered decision support systems often make less accurate decisions on average than AI-powered systems on their own [4, 8, 9, 11, 29, 34, 40, 62]. This is surprising because if people combined their own knowledge with the information provided by the decision support systems, the resulting decisions should be more accurate than those made by either unaided people or AI-powered systems alone. A number of researchers investigated possible reasons for these surprising results. There is converging evidence that people overrely on the AI-generated recommendations [8, 34, 39, 66] and that providing explanations for the AI recommendations might even exacerbate the problem [4]. Recent work demonstrated that certain kinds of in-the-moment interventions—such as forcing people to wait before submitting a decision, having people state their initial decision before seeing the AI recommendation, or having the AI recommendation provided only on demand—can reduce (but not eliminate) this overreliance [9]. It is possible that these interventions are effective because they encourage people to engage more deeply with the AI-provided information. We examine one of these interventions (the one where people state their initial decision before seeing the AI recommendation) in Experiment 2.

Research in cognitive psychology suggested that people process information on different levels. Deep processing occurs when individuals engage in more meaningful analysis of information and link it to existing knowledge structures [2]. In learning sciences, depth of processing is often associated with the degree of cognitive engagement, which is described as a "psychological state in which students put in a lot of effort to truly understand a topic and in which students persist studying over a long period of time." [59].

While some authors discuss cognitive engagement as a personal trait of a student that does not depend on context [3], others suggest that cognitive engagement depends on the structure of each task [15, 30, 59]. For example, searching for information on the Internet or engaging in discussion with other students engenders higher levels of cognitive engagement than passively listening to a lecture and results in higher learning gains. Rotgans and Schmidt attributes these differences in cognitive engagement to different degrees of autonomy afforded by different learning tasks [59]. Chi et

al. propose Interactive-Constructive-Active-Passive (ICAP) framework to describe a continuum of learning behaviors (from passive, to active, to constructive, to interactive) and argue that each subsequent level leads to an increase in cognitive engagement and learning [15, 16]. Further building upon the ICAP framework, Lam and Muldner showed that engaging in collaborative constructive activities has a positive impact on learning [41].

In this study, we build upon these investigations in learning sciences and examine learning gains resulting from engagement with different types of AI-generated information. Previous research proposed a number of instruments for measuring cognitive engagement directly. However, many of the instruments developed thus far take a more longitudinal view of engagement and focus on completion of multiple activities involved in learning (such a rate of completion of homework, etc. as in [3]). These instruments are ill-suited to capture variability in cognitive engagement within each individual task. Rotgans and Schmidt [59] proposed an instrument more sensitive to the degree of cognitive engagement in real time; however, this instrument is still highly tailored to educational settings where learning is the primary goal, rather than brief tasks where learning is incidental. Consequently, in this study we did not use more direct measures of cognitive engagement, and instead we measured how much people learned.

Incidental learning typically occurs as a byproduct of other activities (e.g., problem solving, advice seeking) rather than as a result of explicit or formal educational activities [47]. However, like formal learning, incidental learning can only occur if people engage deeply with information. Incidental learning is common in professional settings [47, 48], and it can result from a wide range of professional activities and interactions, including colleagues assisting each other on decision-making tasks [6].

One recent project already demonstrated that people can learn from AI-generated recommendations and explanations provided that the explanations are carefully crafted to include relevant domain-specific information [18]. However, that project was conducted in the context of chess playing: participants received AI-generated recommendations and explanations on what moves to make. Because of the unique nature of the task (people who play chess typically are cognitively engaged in the game), it is unclear how broadly the results of that study generalize.

More extensive research on incidental learning has been done in an adjacent domain: navigational aids. A number of studies (e.g., [7, 20, 52]) demonstrated that the design of the navigation aid interface can influence both how well people can navigate in the moment and how much they learn about the route and the spatial configuration of the surrounding environment. While some researchers found trade-offs between how well different designs supported navigation and learning [52], Dey et al found an approach that supported both [20]: They contrasted two designs that showed participants their location on the map. In one design, participants were also shown directional arrows telling them when and in what direction to turn, while the other design provided no such additional information. Both designs were equally effective at supporting navigation, but the latter design (the one without directional arrows) resulted in significantly greater learning than the one with the arrows. Generalizing beyond the domain of navigation support, these results suggest that individuals actively engage with and synthesize information when they need to arrive at a decision on their own; when the solution is presented to them as a suggestion, they process information more superficially, which inhibits learning. We build on this insight in Experiment 3.

3 EXPERIMENT 1: AI PROVIDES RECOMMENDATIONS AND EXPLANATIONS

The purpose of the first experiment was to evaluate whether incidental learning occurs when people receive decision support from a simple explainable AI system, one that offers a decision recommendation and an explanation. We hypothesized that such a design would improve people's immediate task performance compared to situations when no AI support was offered, but would not result in learning.

3.1 Tasks and Conditions

To ensure that the results are informative, we used actual decision-making tasks rather than proxy tasks [8]. Specifically, we adopted the nutrition knowledge quiz (following the design of [10]), in which participants were presented with images and descriptions of pairs of meals and were asked which meal contained more of one of the four macronutrients (carbohydrates, fat, protein, or fiber). Questions were designed such that each depended on one particular nutrition concept, such as that avocados are a significant source of fat or that soy beans contain more protein than most other beans.

For each nutrition concept, we prepared three questions. They were used as a pre-test, the intervention, and post-test, respectively.

The first question relating to a particular nutrition concept (the pre-test) was used to measure the participant's pre-existing knowledge of that concept. After responding to the pre-test question, they received only correctness feedback ("Correct" or "Not quite"). Prior work using this task design suggests that simple correctness feedback results in as little learning as no feedback at all [10], while our pilot data indicated that participants preferred receiving some feedback. Thus, providing simple correctness feedback minimized learning from the first exposure to the concept while improving participant experience.

The second time a participant encountered a particular concept (the intervention), they were presented with a design corresponding to one of the three experimental conditions (see the next section). By comparing participants' performance on the second exposure to a concept (and normalizing by their performance on the pre-test; see Section 3.4), we could estimate how much a particular design supported participants in performing the task at hand.

The third question in each concept (the post-test) served as a means to assess how much participants learned about the concept from the intervention. After responding to the post-test, participants were presented both with correctness feedback and with an explanation — this decision had no impact on the data collected and was, instead, intended to improve participants' experience.

The concepts were drawn at random from a larger pool, such that each study included two concepts for each of the four macronutrients (for a total of 8 concepts). Concepts were randomly assigned to conditions. The order in which questions were presented was randomized (thus, for each participant, different questions served

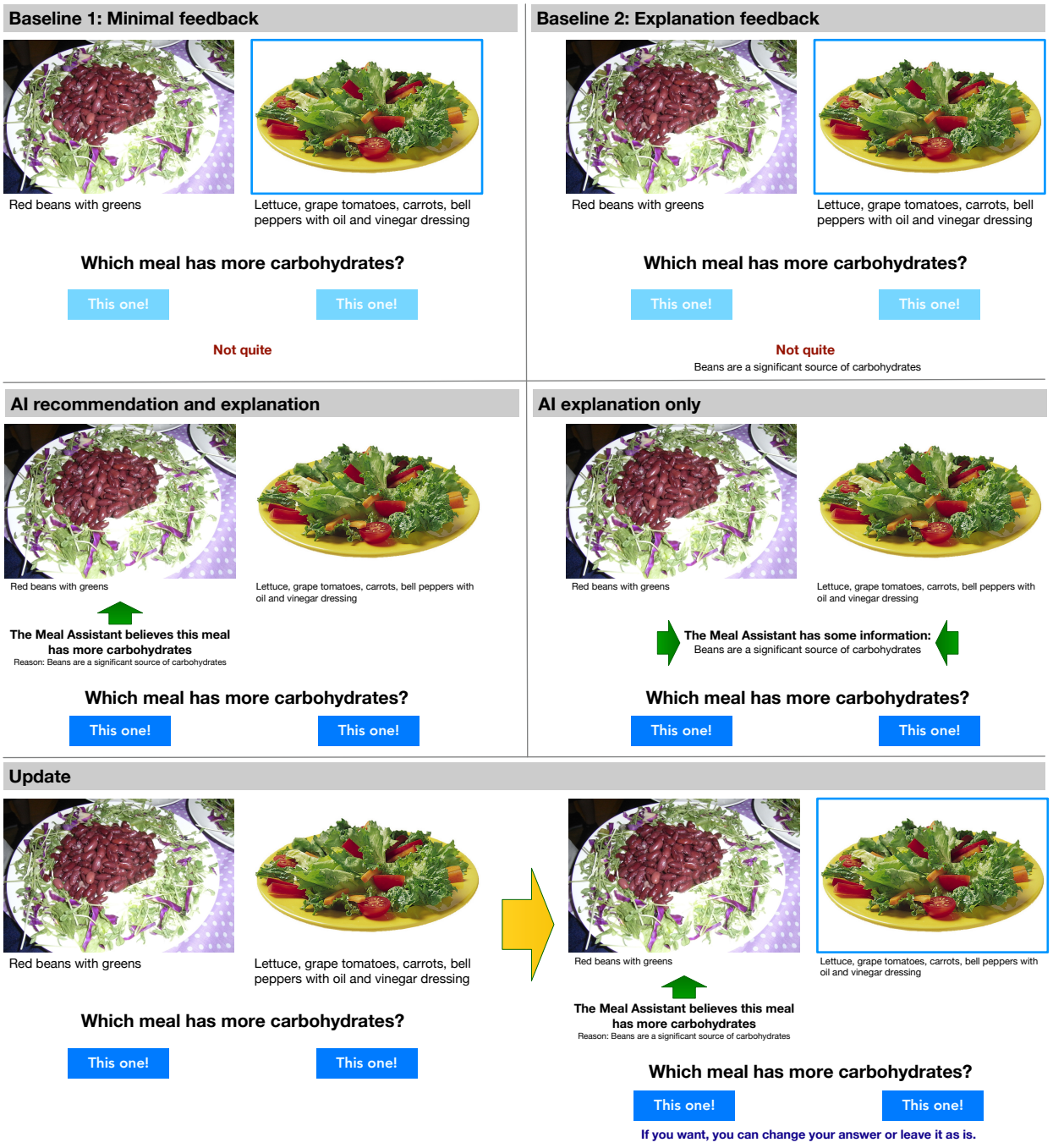

Figure 1: Experimental conditions included in the experiments. Top-left: the *Minimal feedback* design, a non-AI baseline condition where participants only receive correctness feedback ("Correct!" or "Not quite") after they make their decision. Top-right: the *Explanation feedback* design, the second non-AI baseline condition, in which participants receive both correctness feedback and a brief explanation upon making their decision. Middle-left: *AI recommendation and explanation* design (Experiment 1)—participants see a recommendation and an explanation from a simulated AI systems, called the Meal Assistant, prior to making their own decision. Middle-right: *AI explanation only* design (Experiment 3)—participants see only an explanation (but no recommendation) from a simulated AI system before making their own decision. Bottom: *Update* design (Experiment 2)—participants first make their own decision and then they are shown a recommendation and explanation from the Meal Assistant. They have the option to change their decision at this point.

as pre-test, intervention, and post-test). Consequently, the order of questions within each concept was randomized and the concepts were intermixed. There were a total of 24 questions (8 concepts × 3 questions per concept).

This experiment included three conditions: two baselines and a condition simulating a common approach to designing AI-powered decision-support tools:

- **Baseline 1, *Minimal feedback.*** In this condition, which is illustrated in Figure 1 (top left), participants received no AI assistance and only minimal correctness feedback ("Correct" or "Not quite") after they submitted their answer. As mentioned before, in prior research this design was not significantly different from a design where no feedback at all was provided [10]. Thus, we consider this to be the low baseline—it is unlikely that any other condition will result in less learning.
- **Baseline 2, *Explanation feedback.*** In this condition, illustrated in Figure 1 (top right), participants again received no AI assistance, but they received both correctness feedback ("Correct" or "Not quite") and a brief explanation (e.g., "Avocados are a significant source of fat") after they submitted their response. The same explanation was provided whether the participant answered correctly or not. In prior work, such feedback resulted in significantly greater learning than conditions where only correctness feedback was provided or where no feedback was provided [10]. Thus, we consider this to be a high baseline: while an even more effective design might be possible, it represents a demonstrably effective solution.
- ***AI recommendation and explanation.*** This condition (Figure 1 middle left) simulated the way AI-powered decision support systems are frequently implemented today: the AI recommendation accompanied by an explanation was presented at the very moment the person was presented with a decision task. To simulate real decision-making tasks where the ground truth is unknown, no feedback was provided to participants after they made their decision. Instead, they were told "Your response has been recorded. (you will receive feedback at the end of the test)". In the study instructions, the AI was introduced as the "Meal Assistant, which is an experimental computer system that can analyze the nutritional content of meals." Additionally, they were told that "The Meal Assistant is right most of the time but not always. You are welcome to consider its recommendations, but you should make whatever decision you think is best." This uncertainty is typical for contemporary AI-powered decision support, and can lead to different degrees of trust in AI-generated information. However, given that the focus of this study was on cognitive engagement, rather than trust, we designed our study such that the Meal Assistant recommendations were always correct.

The explanations presented in the *Explanation feedback* and in the *AI recommendation and explanation* conditions were identical. They were also designed to take the form of *contrastive* explanations. Contrastive explanations show, for example, why a diagnosis should be disease X *instead of* disease Y, where disease Y, used for contrast, is known as the *foil*. When there are only two possible choices, the foil is implicit. Contrastive explanations include only information about what is relevant for choosing one option over the foil. There is broad consensus that contrastive explanations are among the most effective in human discourse [45, 51, 64]. Also, the prior work that demonstrated that learning can occur in the context of the Nutrition Knowledge Test also used contrastive explanations (generated by an expert nutritionist) [10]. Finally, recent work demonstrated that explanations that included explicit contrasts were more effective at causing people to select healthy meal alternatives than explanations that contained identical information about the meals but without the explicit comparison [53].

3.2 Procedures

Participants were recruited via two mechanisms: LabintheWild.org and Amazon Mechanical Turk (MTurk). LabintheWild is a platform for conducting online experiments with unpaid participants [58]. Instead of being paid, participants are incentivised by the promise that at the end of the study they will see their own results and compare themselves to other test takers. Both curiosity and opportunities for social comparison have been shown to increase engagement of online participants [32, 42] and multiple validation studies demonstrated that data collected on LabintheWild and other similar platforms are valid and lead to the same conclusions as data collected in traditional laboratory settings [26, 31, 43, 44, 58]. While some LabintheWild studies attracted tens of thousands of participants [25, 57], experimenters have little control over the rate at which participants arrive. Thus, we supplemented recruitment with MTurk, which is also an effective choice collecting valid behavioral data [37]. We paid MTurk participants $1 (US) aiming for $10/hour (the median time to complete the study was 6 minutes).

Upon arriving at the experiment web site, all participants were presented with brief information about the study (including a promise to see their own results and the aggregate results of others at the end) followed by informed consent. Next, participants were asked if they had taken this study before and they were presented with a demographics form (where all questions were optional). Following [60], we offered five options for gender: male, female, non-binary, prefer to self-describe (which enabled a free response field), and prefer not to answer. Then, they were presented with the instructions. Next, they completed the main nutrition knowledge test consisting of 24 questions. After the main test, they were asked if they had experienced any interruptions, technical difficulties, or if they cheated in any way. Finally, they were shown their own results, the average accuracy of other test takers, and the correct responses (and explanations) for all the questions in the test.

LabintheWild participants were also given an option to share the study via social media or to explore other studies hosted on LabintheWild. Meanwhile, MTurk participants were given given a verification code to enter back on the MTurk web site.

3.3 Approvals

This experiment (as well as the subsequent experiments reported in this manuscript) was reviewed and approved by the Internal Review Board at Harvard University, protocol number IRB15-2398.

	Experiment 1	Experiment 2	Experiment 3	Experiment 3 replication
n	251	268	221	270
Source	LabintheWild: 217 MTurk: 34	LabintheWild: 184 MTurk: 84	LabintheWild: 36 MTurk: 185	LabintheWild: 300 MTurk: 0
Age	11–90, M=33, SD=14.4	6–75, M=33, SD=14.6	9–82, M=38, SD=12.6	11–81, M=28, SD=14.5
Gender	female: 114 male: 112 non-binary: 4 self-described: 1 not responded: 20	female: 139 male: 100 non-binary: 3 self-described: 1 not responded: 25	female: 98 male: 114 non-binary: 3 self-described: 0 not responded: 6	female: 137 male: 88 non-binary: 12 self-described: 1 not responded: 62

Table 1: Participants

3.4 Design and Analysis

This was a within-subjects experiment with three conditions (*Minimal feedback*, *Explanation feedback*, and *AI recommendation and explanation*).

We collected two dependent measures, separately for each condition:

- **Immediate benefit.** We quantified the improvement at the intervention time compared to pre-test by computing normalized change c [49] (one of the measures commonly used to measure student improvement) between the average correct rate for intervention questions and the average correct rate for the pre-test questions:

$$c = \begin{cases} \frac{\text{intervention} - \text{pre}}{1 - \text{pre}} & \text{if intervention} > \text{pre} \\ \frac{\text{intervention} - \text{pre}}{\text{pre}} & \text{if intervention} < \text{pre} \\ 0 & \text{if intervention} = \text{pre} \end{cases} \quad (1)$$

  where "intervention" stands for the average correct response rate during intervention trials and "pre" denotes the average correct response rate during the pre-test trials.
- **Learning.** To quantify learning, we used an analogous approach to compute normalized change between the average correct rate at the post-test and the average correct rate at the pre-test.

Based on the results from the prior research that measured learning in the context of the Nutrition Knowledge Test [10], we powered our experiment to detect differences corresponding to Cohen's $d \geq 0.2$. Given the within-subjects design of our experiment, this meant a minimum of 199 participants.

Our measures were not normally distributed. Therefore, we used a non-parametric test, Wilcoxon signed-rank test, to test for statistically significant differences in our data.

We computed effect sizes using a standard approach for Wilcoxon non-parametric tests [23, 61]: $r = \frac{Z}{\sqrt{n}}$ where Z is the test statistic produced by the Wilcoxon signed rank test and n is the number of participants in the sample. The interpretation of this effect size follows Cohen's guidelines for r: .5 means a large effect, .3 a medium effect, and .1 is a small effect [17].

3.4.1 Treatment of Outliers. We removed from the analyses all participants who indicated that they had taken the test before. Otherwise, we kept all the participants. However, we did analyze our data for extreme outliers to make sure that they did not improperly affect the results. Because our primary outcome measure was bounded (and extreme values were plausible), we used trial completion times as indicators of unusual behavior. Specifically, we flagged all trials that took more than 1 minute to complete and we repeated our analyses after removing all participants who had at least one outlier trial. All the conclusions in all the experiments still held after outliers were removed.

3.5 Results

3.5.1 Participants. 251 people participated in this experiment. Their demographics are summarized in Table 1.

3.5.2 Main Results. The key results are visualized in Figure 2.

As hypothesized, participants demonstrated a significantly higher immediate benefit of the intervention in the *AI recommendation and explanation* condition (normalized change between intervention and pre-test: M=0.341) than in either *Explanation feedback* condition (M=0.158, Z=3.55, p=0.0004, r=0.22) or the *Minimal feedback* condition (M=0.167, Z=3.56, p=0.0004, r=0.22). There was no significant difference in in terms of the immediate benefit between the *Explanation feedback* and *Minimal feedback* conditions (Z=0.10, n.s.).

The *Explanation feedback* condition resulted in significantly larger learning gain (normalized change between post-test and pre-test: M=0.325) than the *AI recommendation and explanation* condition (M=0.206, Z=2.37, p=0.0179, r=0.15) or the *Minimal feedback* condition (M=0.189, Z=3.14, p=0.0017, r=0.20). Consistent with our hypothesis, there was no significant difference in learning gain between the *AI recommendation and explanation* and the *Minimal feedback* conditions (Z=0.68, n.s.).

4 EXPERIMENT 2: PARTICIPANTS MAKE INITIAL DECISIONS BEFORE SEEING AI RECOMMENDATIONS AND EXPLANATIONS

In this experiment, we tested if people experience the benefit of incidental learning if they first make their own decision, and only

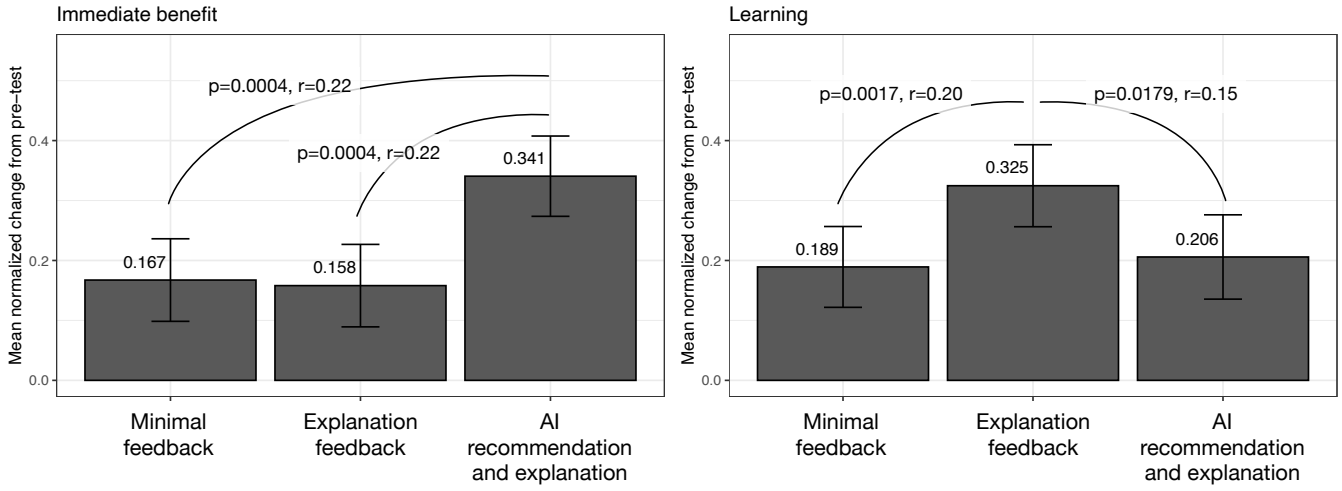

Figure 2: Experiment 1 results. Left: the immediate benefit per condition. Right: learning per condition. All results are reported as mean normalized changes from the pre-test. Error bars show 95% confidence intervals.

then are presented with the AI recommendation and explanation (and an option to revise their initial decision). We refer to this form of human-AI interaction as the *Update* design. In prior work the *Update* design resulted in more accurate decisions [22, 29] and reduced overreliance on the AI [9] compared to the *AI recommendation and explanation* design. Buçinca et al [9] hypothesized that the *Update* design induces people to engage cognitively with the AI-provided information. Thus, we hypothesized that the *Update* design would result both in improved task performance and learning.

4.1 Tasks and Conditions

We used the same task design as in Experiment 1. In this experiment, we had the following conditions:

- **Baseline 1, *Minimal feedback.*** Just like in Experiment 1.
- **Baseline 2, *Explanation feedback.*** Just like in Experiment 1.
- ***Update*** This condition is illustrated in Figure 1 (bottom) and follows the general design used in prior research [9, 22, 29]. In this condition, participants first made a decision on their own and only then they were presented with the Meal Assistant recommendation and explanation. At this point they could, but did not have to, change their answer. The Meal Assistant recommendation and explanation was presented regardless of whether the participant's answer was correct. Participants did not know on which tasks they would receive the Meal Assistant recommendation after providing their initial answer.

4.2 Procedures, Design and Analysis

The procedures, experiment design, measures, and analysis approach were the same as in Experiment 1 with one exception: immediately after the main test and before asking if they experienced any interruptions, we asked participants to fill out an abbreviated Need for Cognition questionnaire. Following [24], we used a four-item subset of a common 18-item instrument [13]. We elaborate the reasons for including this measure in Section 5.4.

4.3 Results

4.3.1 Participants. 268 people participated in this experiment. Their demographics are summarized in Table 1.

4.3.2 Main Results. The key results are visualized in Figure 3.

As hypothesized, participants experienced significantly higher immediate benefit in the *Update* condition (normalized change compared to pre-test M=0.391) compared to either the *Minimal feedback* (M=0.110, Z=5.03, p<0.0001, r=0.31) or *Explanation feedback* conditions (M=0.178, Z=3.67, p=0.0002, r=0.22). As expected, this improvement was due to participants changing their answers in response to the Meal Assistant's recommendations—their final answers were significantly more correct than their initial ones (M=0.129, Z=8.17, p<0.0001, r=0.50). There were no statistically significant differences in terms of the immediate benefit among the answers in the *Minimal feedback* condition, the *Explanation feedback* condition, or the initial answers in the *Update* condition.

Contrary to our expectations, participants did not learn more in the *Update* condition (M=0.121) than in the *Minimal feedback* condition (M=0.157, Z=0.71, n.s., r=0.04). Participants learned significantly more in the *Explanation feedback* condition (M=0.354) than in either *Update* (Z=4.59, p<0.0001, r=0.28) or *Minimal feedback* (Z=3.76, p=0.0002, r=0.23) conditions.

5 EXPERIMENT 3: AI PROVIDES EXPLANATIONS ONLY

In this experiment, we evaluated the *AI explanation only* design (Figure 1 middle right), in which people were presented just with the AI-generated explanation (e.g., "Avocados are a significant source of fat") but no explicit decision recommendation. As discussed in Section 2, this design was informed by the prior research on

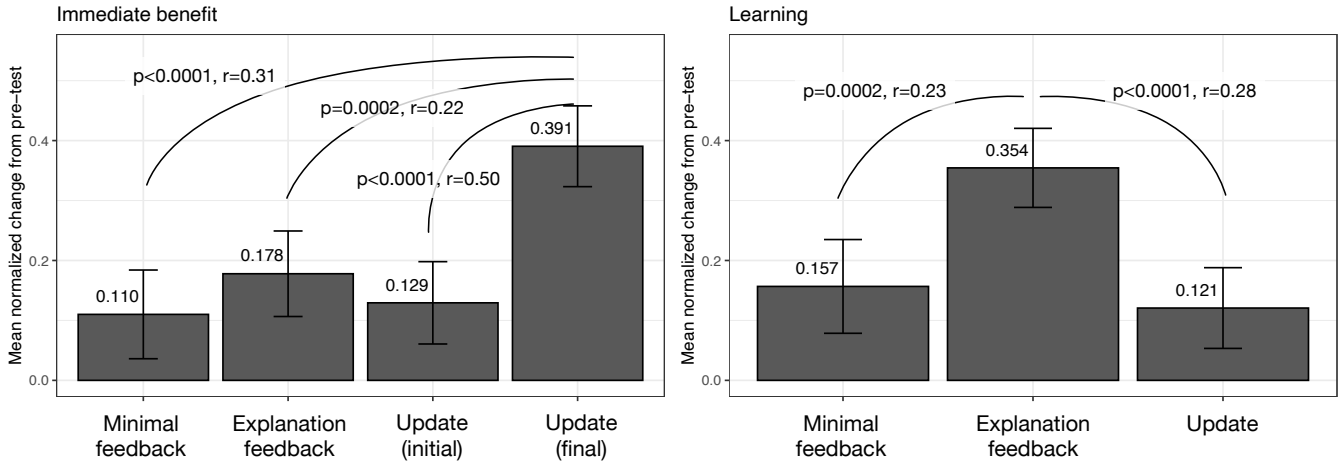

Figure 3: Experiment 2 results. Left: the immediate benefit per condition. Right: learning per condition. All results are reported as mean normalized changes from the pre-test (see Section 3.4 for the definition of normalized change). Error bars show 95% confidence intervals.

navigation aids, which suggested that when people are presented with the necessary information to make a decision but not with an explicit decision suggestion, they actively process the provided information resulting in good performance on the task at hand and in incidental learning [20]. Consequently, we hypothesized that the *AI explanation only* design would result both in improved task performance and learning.

5.1 Tasks and Conditions

We used the same task design as in Experiments 1 and 2. In this experiment, we had the following conditions:

- **Baseline 1, *Minimal feedback.*** Just like in Experiments 1 and 2.
- **Baseline 2, *Explanation feedback.*** Just like in Experiments 1 and 2.
- ***AI explanation only.*** This condition is illustrated in Figure 1 (middle right). Participants were presented with an explanation, but no recommendation as to which answer was correct. For example, participants were told that milk is a significant source of carbohydrates, but they had to process this information themselves to decide which answer to select. As in the *AI recommendation and explanation* and *Update* conditions, participants were not provided with any feedback on their answers on trial during which AI assistance was offered.

5.2 Procedures, Design and Analysis

All methods were the same as in Experiment 2 with one addition: after we conducted and analyzed the data from the experiment, we replicated it on an new independent sample.

5.3 Results

5.3.1 Participants. 221 people participated in the initial experiment and 270 participated in the replication. Their demographics are summarized in Table 1.

5.3.2 Main Results. The key results are visualized in Figure 4.

As hypothesized, participants experienced greater immediate benefit in the *AI explanation only* condition (M=0.422) than in either *Minimal feedback* (M=0.158, Z=4.54, p<0.0001, r=0.31) or *Explanation feedback* (M=0.144, Z=4.84, p<0.0001, r=0.33) conditions.

Also as expected, participants learned significantly more in the *AI explanation only* condition (M=0.342) than in the *Minimal feedback* condition (M=0.138, Z=3.39, p=0.0007, r=0.23). As before, participants also learned more in the *Explanation feedback* condition (M=0.320) than in the *Minimal feedback* condition (Z=3.28, p=0.0010, r=0.22). There was no statistically significant difference between *Explanation feedback* and *AI explanation only* conditions in terms of learning (Z=0.41, n.s., r=0.03).

The results of the replication (summarized in Figure 5) supported all of the conclusions from the initial experiment: all the significant differences remained significant with comparable effect sizes.

5.4 Audit for Intervention Generated Inequalities

A design intervention, even if it is helpful on average, can be more useful to some groups than others resulting in *intervention-generated inequalities* [65]. This is especially problematic if the intervention benefits an already privileged group more than others. An internal audit [56], which involves disagreggating the results by relevant demographic factors, can help uncover such problems. Given that our intervention targets the person's motivation to exert cognitive effort to engage with the AI-generated information, we built on similar prior work [9] and disaggregated our results by Need for Cognition (NFC), a stable personality trait that reflects

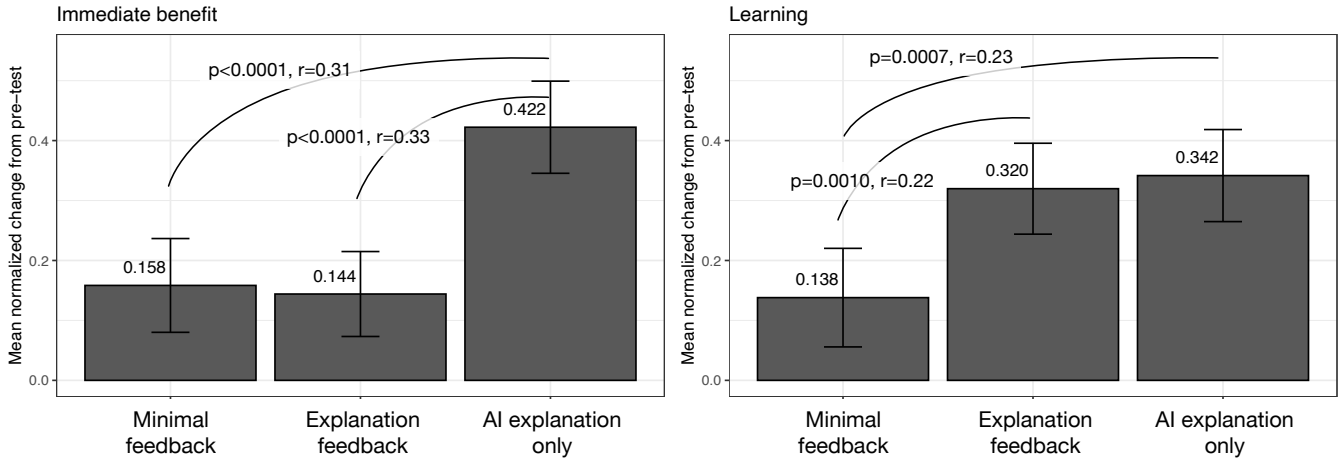

Figure 4: Experiment 3 results. Left: the immediate benefit per condition. Right: learning per condition. All results are reported as mean normalized changes from the pre-test (see Section 3.4 for the definition of normalized change). Error bars show 95% confidence intervals.

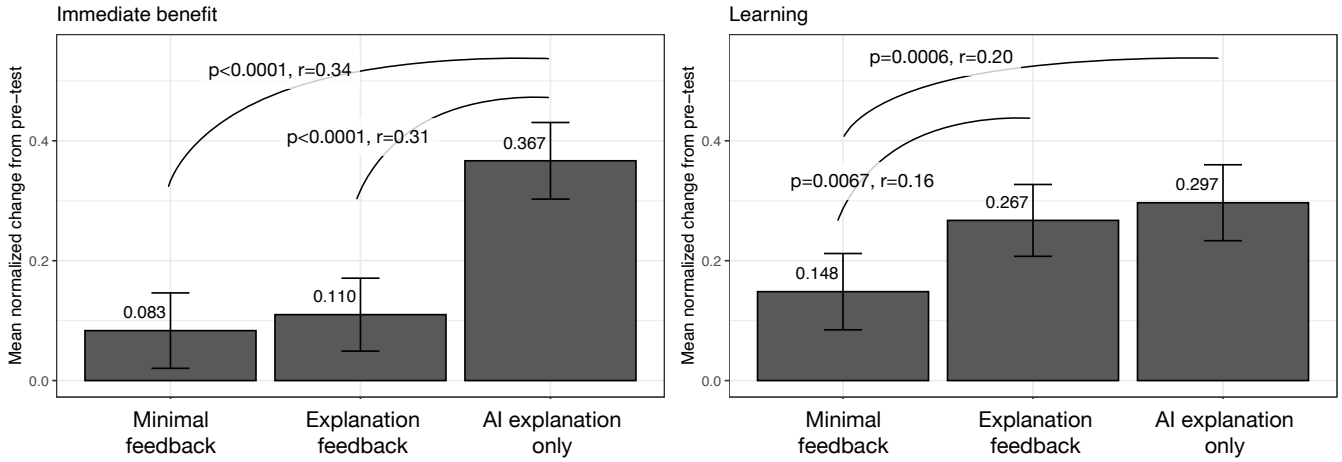

Figure 5: Experiment 3 replication results. Left: the immediate benefit per condition. Right: learning per condition. All results are reported as mean normalized changes from the pre-test. Error bars show 95% confidence intervals.

how much a person enjoys engaging in effortful cognitive activities [12, 55].

For this analysis, we used the data from both the original experiment and the replication. With NFC measured on a 1–5 scale, we divided our participants into two groups of roughly equal size: the Low NFC group (NFC $\leq$ 3.25; n=229) and the High NFC group (NFC $>$ 3.25; n=207).

The results of our audit are illustrated in Figure 6. We did not observe statistically significant differences in learning between the two NFC groups in the *Minimal feedback* condition (Wilcoxon rank sum test Z=0.668, n.s.) or in the *Explanation feedback* condition (Z=0.677, n.s.). However, in the *AI explanation only* condition the High NFC group learned significantly more (M=0.388) than the Low NFC group (M=0.255, Z=2.35, p=0.02). This result indicates that our intervention might have disparate effects for different individuals depending on their level of cognitive motivation.

6 DISCUSSION AND CONCLUSION

In this work, we examined how individuals engage with different types of AI-generated assistance and the impact of AI support on task performance and incidental learning. Specifically, we focused on the impact of decision recommendations and explanations, two increasingly popular components of AI-generated support with growing availability in both professional decision support tools [5] and personal informatics systems [19]. Previous research already demonstrated the positive impact of explanations provided as feedback on the task on incidental learning [10]; as a result, we expected that explanations would be critical to learning. Furthermore, we were interested in examining interconnections between decision

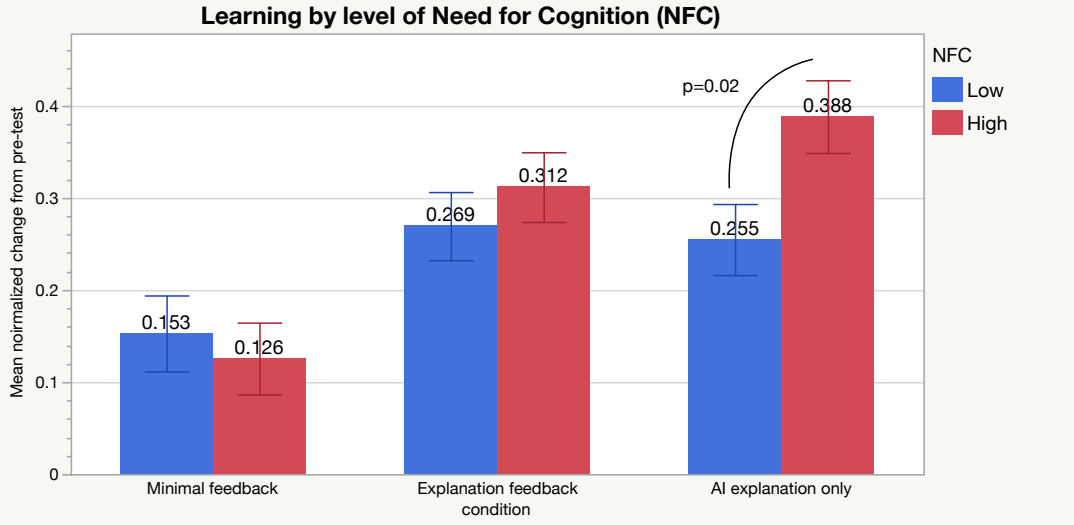

Figure 6: Results from Experiment 3 disaggregated by Need for Cognition. Error bars show 95% confidence intervals.

recommendations and explanations and their individual and combined impact on incidental learning.

Our results showed that, as expected, all designs of the human-AI interactions we have tested provided significant immediate benefit, helping people make better decisions in those situations in which the AI assistance was offered. These results contribute to the growing body of evidence showing that people working with AI-powered decision support tools often (but not always [34, 62]) make more accurate decisions than they would have on their own [11, 29, 40].

As hypothesized, we observed the evidence of incidental learning in the *AI explanation only* condition but not in the *AI recommendation and explanation* condition. As shown in Figure 1, in the *AI recommendation and explanation* condition participants were presented with both a decision recommendation and an explanation, whereas in the *AI explanation only* condition they received just an explanation — if they wanted to benefit from it, they had to process it carefully enough to infer which decision the explanation supported. While prior work has highlighted the critical role of explanations in promoting learning [10, 18], our work additionally demonstrated the value of creating the conditions for learners to engage constructively (as defined in the ICAP framework [15, 16]) with the explanations.

Contrary to our hypothesis, we did not observe any evidence of learning in the *Update* design, as compared to the *Explanation feedback* baseline. There are multiple possible explanations to this unexpected finding. First, when comparing this condition with the *Explanation feedback* baseline, it is possible that differences in the attributed source of information between human experts (*Explanation feedback* baseline) and AI (*Update* design) led participants to place different emphasis on otherwise identical information and engage with information provided by human experts deeper than with information provided by an AI. It is also possible that this lack of learning could be attributed to the framing of this information as either direct task feedback (*Explanation feedback* baseline) or as additional information for contemplation (*Update* design) with people examining the task feedback more carefully than the optional additional information. Furthermore, it is possible that even though the participants were given a chance to update their decisions based on additional AI-generated information in the *Update* condition, they did not fully engage in the synthesis and simply accepted or rejected the recommended answer. This would suggest that this design did not fully reach the constructive level from the ICAP framework [15, 16]. Although this design has been previously shown to reduce overreliance on the AI recommendations [9], perhaps this effect was achieved by people reflecting more deeply on their own knowledge (and being more likely to follow their own judgement) rather than by carefully combining the AI-provided information with their own knowledge.

The finding that the *AI explanation only* design appears to benefit people with high Need for Cognition (NFC) more than those with low NFC adds to the growing body of evidence suggesting that adding any form of "intellligence" to interactive systems generally benefits high NFC individuals the most [9, 14, 24, 27]. It is a potential source of concern because it suggests that the contemporary trends in interactive computing may be creating disparities that had not existed before.

A key limitation of our work is that our simulated AI (the Meal Assistant) was always correct. This is a strength in the context of Experiments 1 and 2 as it shows that even in the idealized conditions the *AI recommendation and explanation* and *Update* designs did not support incidental learning. However, further work is needed to determine how sensitive the *AI explanation only* design is to AI errors, and whether information on certainty of AI-recommendations can have an impact on engagement with explanations. A complementary future direction is to examine if the *AI explanation only* design helps to reduce human overreliance on the AI. Another possible limitation is that our study included low-risk, inconsequential decisions and individuals had no special expertise in nutrition and nutritional judgment. Thus, these findings may not generalize to AI-assisted decision making by experts in domains with more consequential

decisions; it is plausible that in those contexts individuals may exhibit deeper engagement with AI than was captured in our study. Furthermore, we only measured immediate learning but not long-term learning. While our results already indicate likely differences in the level of cognitive engagement across the three experiments, further work is needed to understand the long-term effectiveness of incidental learning from both expert and AI explanations. Lastly, we note that in the *AI explanation only* condition the visual design emphasized the AI-generated explanations (with the green arrows pointing to the explanation) while in the other AI conditions the green arrows pointed to the correct answer. It is possible that by making the explanations more salient, the visual design of the *AI explanation only* condition drew participants' attention to those explanations more than in the other AI conditions. However, we also note that explanations provided in the *Explanation feedback* baseline robustly led to learning even though they were not visually salient. Thus, we are confident that the impact of the visual design in Experiment 3 was very small compared to the impact of the overall design of that condition.

An important consideration in our work is that we used contrastive explanations [45, 64] and there is some evidence that the *AI explanation only* design may be less effective at supporting the task at hand when non-contrastive explanations are used [40]. As previously explained in Section 3.1, contrastive explanations answer the question why choose X instead of Y, where Y is known as the foil. Contrastive explanations are the dominant form of explanations in human-human discourse. They are much briefer than explanations that enumerate all evidence and thus make more efficient use of the cognitive resources of the person receiving the explanation. There is growing recognition that explainable AI systems should provide contrastive explanations [33, 51], but this is not yet the norm. Recent work has produced several methods for efficiently computing contrastive explanations once the foil is known [1, 21, 38, 46, 50, 63]. However, there remains the challenge of identifying a good foil. In some situations—when there are only two options to choose from or when the person reveals their initial choice—the foil is obvious. However, if we were to use the *AI explanation only* design in a setting where many options are available (e.g., medical treatment selection [34]) new methods are needed to predict what options the user is likely to be considering so that the explanation provided addresses the right contrast.

Lastly, we note that there can be a trade off between encouraging deeper cognitive engagement and the efficiency or perceived usability of the different human-AI interaction designs. While it may be appropriate to design for deeper engagement in domains where the cost of errors is high and human expertise is readily available, such designs may incur higher cognitive burden and slower task completion. Future research need to explore trade-offs between different approaches to the design of AI-powered decision support.

Our findings have important implications for the design of AI-powered decision support aids. First, they contribute compelling evidence in support of the emerging concern that people do not carefully process the AI-generated information when given a decision recommendation and an explanation [4, 9, 28]. This, in turn, suggests an urgent need for research on fundamentally novel approaches to human-AI interaction.

Second, our results suggest one such novel approach that places emphasis on synthesizing information necessary to arrive at decisions, but leaving the actual decisions up to human users. Our study provided strong evidence that this approach can promote deeper cognitive engagement with AI-generated information and can lead to higher quality decisions and learning. However, future research should examine trade-offs between cognitive burden associated with deeper engagement and task efficiency, whether similar approaches would hold in higher stakes domains, and the impact of uncertainty in the accuracy of AI-generated information on the degree of cognitive engagement.

ACKNOWLEDGMENTS

We thank Zana Buçinca, Maja Malaya, Barbara J. Grosz, Jean J. Huang, and the members of the Intelligent Interactive Systems group at Harvard for valuable feedback. This work was funded in part by the National Science Foundation (grant IIS-2107391).

DATA AVAILABILITY

Data used in this research can be accessed at https://doi.org/10.7910/DVN/JQL0JW.

REFERENCES

[1] David Alvarez-Melis, Hal Daumé III, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Weight of evidence as a basis for human-oriented explanations. *arXiv preprint arXiv:1910.13503* (2019).

[2] John R Anderson and Lynne M Reder. 1979. An elaborative processing explanation of depth of processing. *L.; S. Cermak and FIM Craik, Eds., Levels of Processing in Human Memory (Erlbam, 1979)* (1979), 385–404.

[3] James J Appleton, Sandra L Christenson, Dongjin Kim, and Amy L Reschly. 2006. Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of school psychology* 44, 5 (2006), 427–445.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proceedings of CHI '21*.

[5] David W Bates, David Levine, Ania Syrowatka, Masha Kuznetsova, Kelly Jean Thomas Craig, Angela Rui, Gretchen Purcell Jackson, and Kyu Rhee. 2021. The potential of artificial intelligence to improve patient safety: a scoping review. *NPJ digital medicine* 4, 1 (2021), 1–8.

[6] Lucy M Berlin and Robin Jeffries. 1992. Consultants and apprentices: observations about learning and collaborative problem solving. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work*. 130–137.

[7] Sven Bertel, Thomas Dressel, Tom Kohlberg, and Vanessa von Jan. 2017. Spatial Knowledge Acquired from Pedestrian Urban Navigation Systems. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 32, 6 pages. https://doi.org/10.1145/3098279.3098543

[8] 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 (IUI '20)*. ACM, New York, NY, USA.

[9] Zana Buccinca, 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

[10] Marissa Burgermaster, Krzysztof Z. Gajos, Patricia Davidson, and Lena Mamykina. 2017. The Role of Explanations in Casual Observational Learning About Nutrition. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 4097–4145. https://doi.org/10.1145/3025453.3025874

[11] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.

[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] Giuseppe Carenini. 2001. An Analysis of the Influence of Need for Cognition on Dynamic Queries Usage. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems* (Seattle, Washington) *(CHI EA '01)*. ACM, New York, NY, USA, 383–384. https://doi.org/10.1145/634067.634293

[15] Michelene TH Chi and Ruth Wylie. 2014. The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist* 49, 4 (2014), 219–243.

[16] Michelene T H Chi. 2009. Active-Constructive-Interactive: A Conceptual Framework for Differentiating Learning Activities. *Topics in Cognitive Science* 1, 1 (2009), 73–105. https://doi.org/10.1111/j.1756-8765.2008.01005.x

[17] Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences* (2nd edition ed.). Lawrence Erlbaum Associates.

[18] Devleena Das and Sonia Chernova. 2020. Leveraging Rationales to Improve Human Task Performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 510–518. https://doi.org/10.1145/3377325.3377512

[19] Pooja M. Desai, Elliot G. Mitchell, Maria L. Hwang, Matthew E. Levine, David J. Albers, and Lena Mamykina. 2019. Personal Health Oracle: Explorations of Personalized Predictions in Diabetes Self-Management. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300600

[20] Sanorita Dey, Karrie Karahalios, and Wai-Tat Fu. 2018. Getting There and Beyond: Incidental Learning of Spatial Knowledge with Turn-by-Turn Directions and Location Updates in Navigation Interfaces. In *Proceedings of the Symposium on Spatial User Interaction* (Berlin, Germany) *(SUI '18)*. Association for Computing Machinery, New York, NY, USA, 100–110. https://doi.org/10.1145/3267782.3267783

[21] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. 2018. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*. 592–603.

[22] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. *Proc. ACM Hum.-Comput. Interact* 5, CSCW2 (October 2021). https://doi.org/10.1145/3479572 arXiv:2109.01443

[23] Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General* 141, 1 (2012), 2.

[24] 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

[25] Krzysztof Z. Gajos, Katharina Reinecke, Mary Donovan, Christopher D. Stephen, Albert Y. Hung, Jeremy D. Schmahmann, and Anoopum S. Gupta. 2020. Computer Mouse Use Captures Ataxia and Parkinsonism, Enabling Accurate Measurement and Detection. *Movement Disorders* 35 (February 2020), 354–358. Issue 2. https://doi.org/10.1002/mds.27915

[26] Laura Germine, Ken Nakayama, Bradley C Duchaine, Christopher F Chabris, Garga Chatterjee, and Jeremy B Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review* 19, 5 (2012), 847–857.

[27] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, Rachel Bellamy, and Klaus Mueller. 2021. Explainable Active Learning (XAL): Toward AI Explanations as Interfaces for Machine Teachers. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 235 (2021), 28 pages. https://doi.org/10.1145/3432934

[28] Ben Green. 2021. The Flaws of Policies Requiring Human Oversight of Government Algorithms. *Available at SSRN* (2021).

[29] 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.

[30] Barbara A Greene, Raymond B Miller, H Michael Crowson, Bryan L Duke, and Kristine L Akey. 2004. Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary educational psychology* 29, 4 (2004), 462–482.

[31] Bernd Huber and Krzysztof Z Gajos. 2020. Conducting online virtual environment experiments with uncompensated, unsupervised samples. *Plos one* 15, 1 (2020), e0227629.

[32] Bernd Huber, Katharina Reinecke, and Krzysztof Z. Gajos. 2017. The Effect of Performance Feedback on Social Media Sharing at Volunteer-Based Online Experiment Platforms. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. ACM, New York, NY, USA, 1882–1886. https://doi.org/10.1145/3025453.3025553

[33] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

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

[35] Ece Kamar. 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence.. In *IJCAI*. 4070–4073.

[36] Ece Kamar, Severin Hacker, and Eric Horvitz. 2012. Combining human and machine intelligence in large-scale crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*. International Foundation for Autonomous Agents and Multiagent Systems, 467–474.

[37] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing performance evaluations of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13)*. ACM, New York, NY, USA, 207–216. https://doi.org/10.1145/2470654.2470684

[38] Benjamin Krarup, Michael Cashmore, Daniele Magazzeni, and Tim Miller. 2019. Model-based contrastive explanations for explainable planning. In *ICAPS 2019 Workshop on Explainable AI Planning (XAIP)*.

[39] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' Deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376873

[40] 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*. 29–38.

[41] Rachel Lam and Kasia Muldner. 2017. Manipulating cognitive engagement in preparation-to-collaborate tasks and the effects on learning. *Learning and Instruction* 52 (2017), 90–101.

[42] Edith Law, Ming Yin, Joslin Goh, Kevin Chen, Michael A. Terry, and Krzysztof Z. Gajos. 2016. Curiosity Killed the Cat, but Makes Crowdwork Better. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (Santa Clara, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 4098–4110. https://doi.org/10.1145/2858036.2858144

[43] Qisheng Li, Krzysztof Z. Gajos, and Katharina Reinecke. 2018. Volunteer-Based Online Studies With Older Adults and People with Disabilities. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. ACM, New York, NY, USA, 229–241. https://doi.org/10.1145/3234695.3236360

[44] Qisheng Li, Sung Jun Joo, Jason D Yeatman, and Katharina Reinecke. 2020. Controlling for participants' Viewing Distance in Large-Scale, psychophysical online experiments Using a Virtual chinrest. *Scientific Reports* 10, 1 (2020), 1–11.

[45] Peter Lipton. 1990. Contrastive explanation. *Royal Institute of Philosophy Supplement* 27 (1990), 247–266.

[46] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable Reinforcement Learning Through a Causal Lens. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*. https://arxiv.org/abs/1905.10958

[47] Victoria J. Marsick and Karen E. Watkins. 2001. Informal and Incidental Learning. *New Directions for Adult and Continuing Education* 2001, 89 (2001), 25. https://doi.org/10.1002/ace.5

[48] Victoria J Marsick, Karen E Watkins, Ellen Scully-Russ, and Aliki Nicolaides. 2017. Rethinking informal and incidental learning in terms of complexity and the social context. *Journal of Adult Learning, Knowledge and Innovation* 1, 1 (2017), 27–34.

[49] Jeffrey D Marx and Karen Cummings. 2007. Normalized change. *American Journal of Physics* 75, 1 (2007), 87–91.

[50] Tim Miller. 2018. Contrastive explanation: A structural-model approach. *arXiv preprint arXiv:1811.03163* (2018).

[51] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269

[52] Stefan Münzer, Hubert D Zimmer, and Jörg Baus. 2012. Navigation assistance: A trade-off between wayfinding support and configural learning support. *Journal of experimental psychology: applied* 18, 1 (2012), 18.

[53] Cataldo Musto, Alain D. Starke, Christoph Trattner, Amon Rapp, and Giovanni Semeraro. 2021. Exploring the Effects of Natural Language Justifications in Food Recommender Systems. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. Association for Computing Machinery, New York, NY, USA, 147–157. https://doi.org/10.1145/3450613.3456827

[54] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 102 (Nov. 2019), 15 pages. https://doi.org/10.1145/3359204

[55] Richard E. Petty and John T. Cacioppo. 1986. The Elaboration Likelihood Model of Persuasion. *Communication and Persuasion* 19 (1986), 1–24. https://doi.org/10.1007/978-1-4612-4964-1_1 arXiv:arXiv:1011.1669v3

[56] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873

[57] Katharina Reinecke and Krzysztof Z. Gajos. 2014. Quantifying Visual Preferences Around the World. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. ACM, New York, NY, USA, 11–20. https://doi.org/10.1145/2556288.2557052

[58] Katharina Reinecke and Krzysztof Z. Gajos. 2015. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Vancouver, BC, Canada) *(CSCW '15)*. ACM, New York, NY, USA, 1364–1378. https://doi.org/10.1145/2675133.2675246

[59] Jerome I Rotgans and Henk G Schmidt. 2011. Cognitive engagement in the problem-based learning classroom. *Advances in health sciences education* 16, 4 (2011), 465–479.

[60] Katta Spiel, Oliver L. Haimson, and Danielle Lottridge. 2019. How to Do Better with Gender on Surveys: A Guide for HCI Researchers. *Interactions* 26, 4 (June 2019), 62–65. https://doi.org/10.1145/3338283

[61] Maciej Tomczak and Ewa Tomczak. 2014. The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in sport sciences* 1, 21 (2014), 19–25.

[62] Michelle Vaccaro and Jim Waldo. 2019. The effects of mixing machine learning and human judgment. *Commun. ACM* 62, 11 (2019), 104–110.

[63] Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, and Mark Neerincx. 2018. Contrastive explanations with local foil trees. *arXiv preprint arXiv:1806.07470* (2018).

[64] Bas Van Fraassen. 1988. The pragmatic theory of explanation. *Theories of explanation* 8 (1988), 135–155.

[65] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (2018), 1080–1088.

[66] 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